MACHINE LEARNING IN DRUG DESIGN

C. DAVID SHERRILL

SCHOOL OF CHEMISTRY AND BIOCHEMISTRY,

SCHOOL OF COMPUTATIONAL SCIENCE AND ENGINEERING

Georgia Tech Institute for Data Engineering and Science

Research Areas:

- Machine Learning
- High-Performance Computing
- Algorithms and Optimization
- Health and Life Sciences
- Materials and Manufacturing
- Energy Infrastructure
- Smart Cities

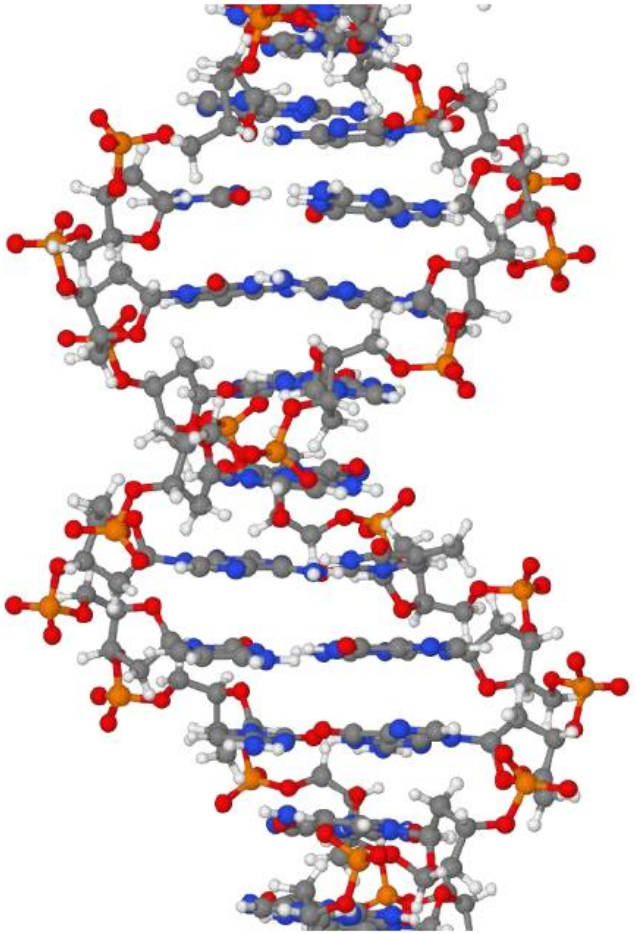
South Big Data Hub:

- NSF-Funded Regional Data Center serving 16 states
- Uses Southern Crossroads, one of the fastest internet gateways in the Southeast

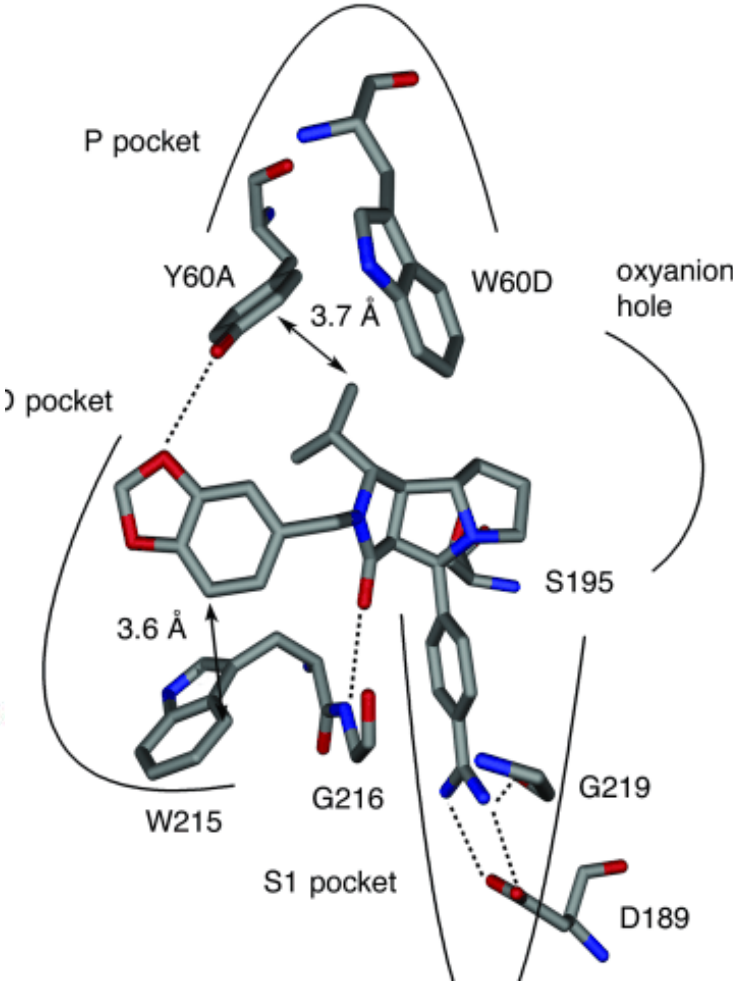
Transdisciplinary Research Institute Advancing Data Science (TRIPODS):

- NSF-funded project involving 39 Georgia Tech faculty
- Scalable inferential strategies, large dataset modeling

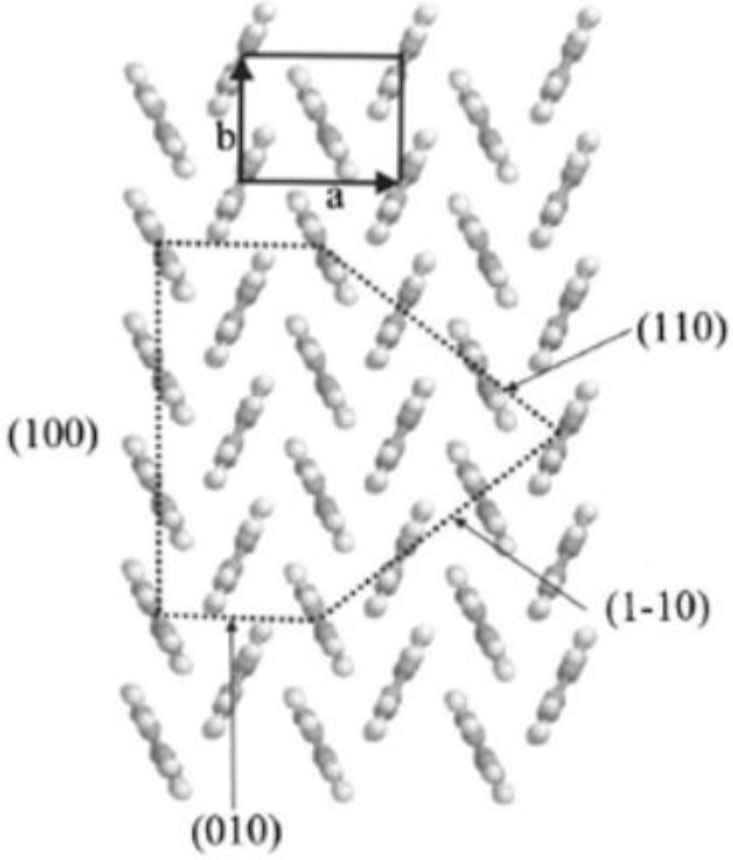
NONCOVALENT INTERACTIONS



Biomolecular structure

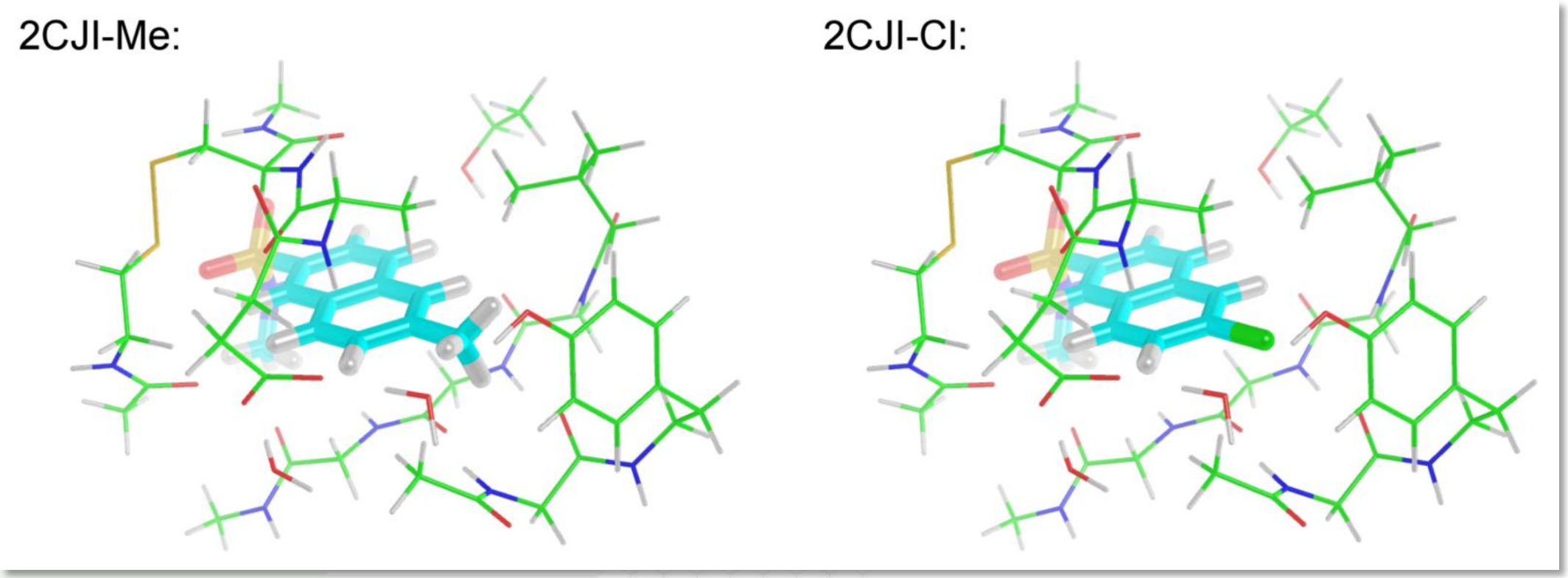


Drug binding

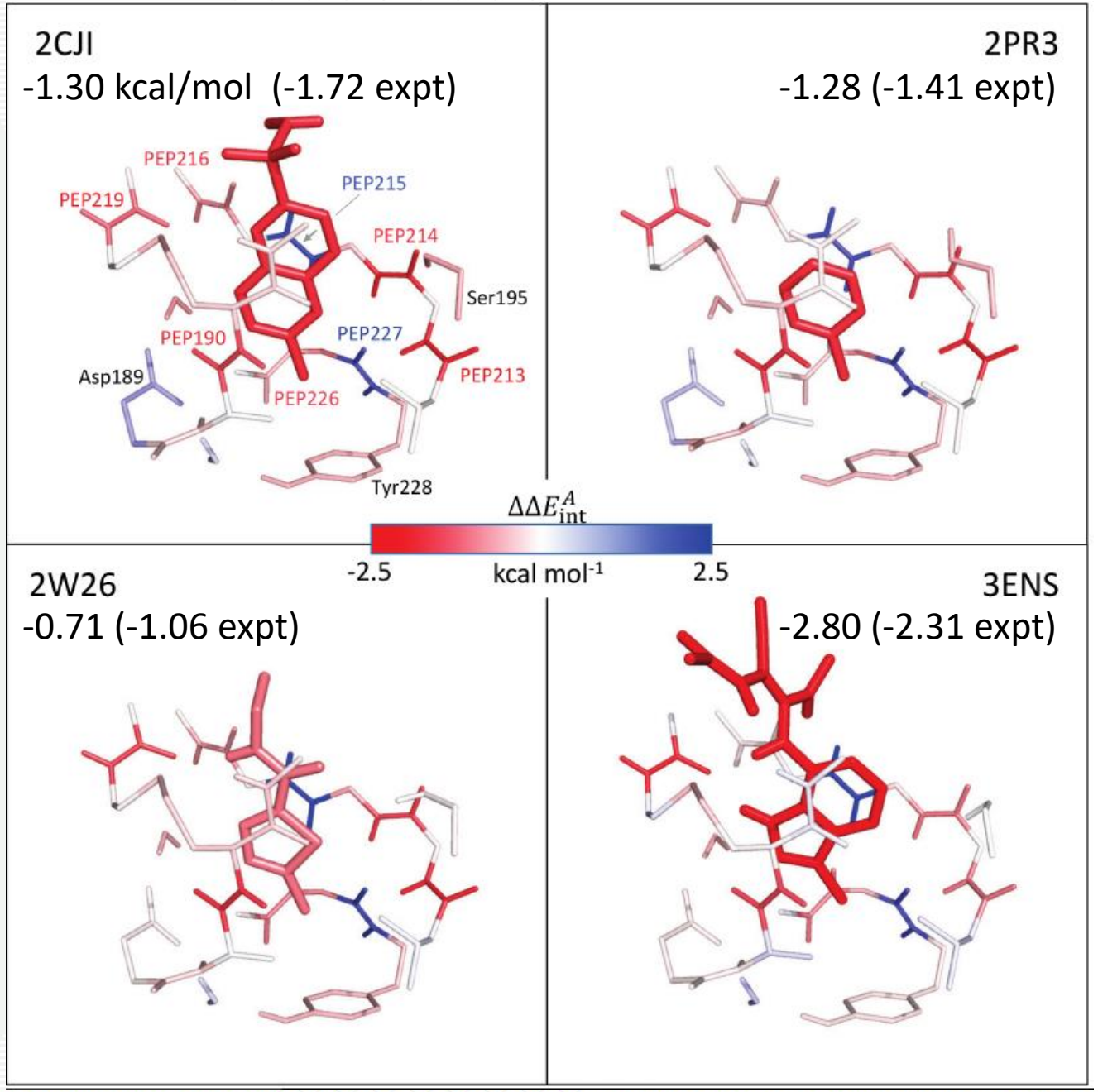


Organic electronics

DRUGS BINDING TO FACTOR XA



The drug with a chlorine atom binds better than the one with a methyl (CH_3) group. Why?

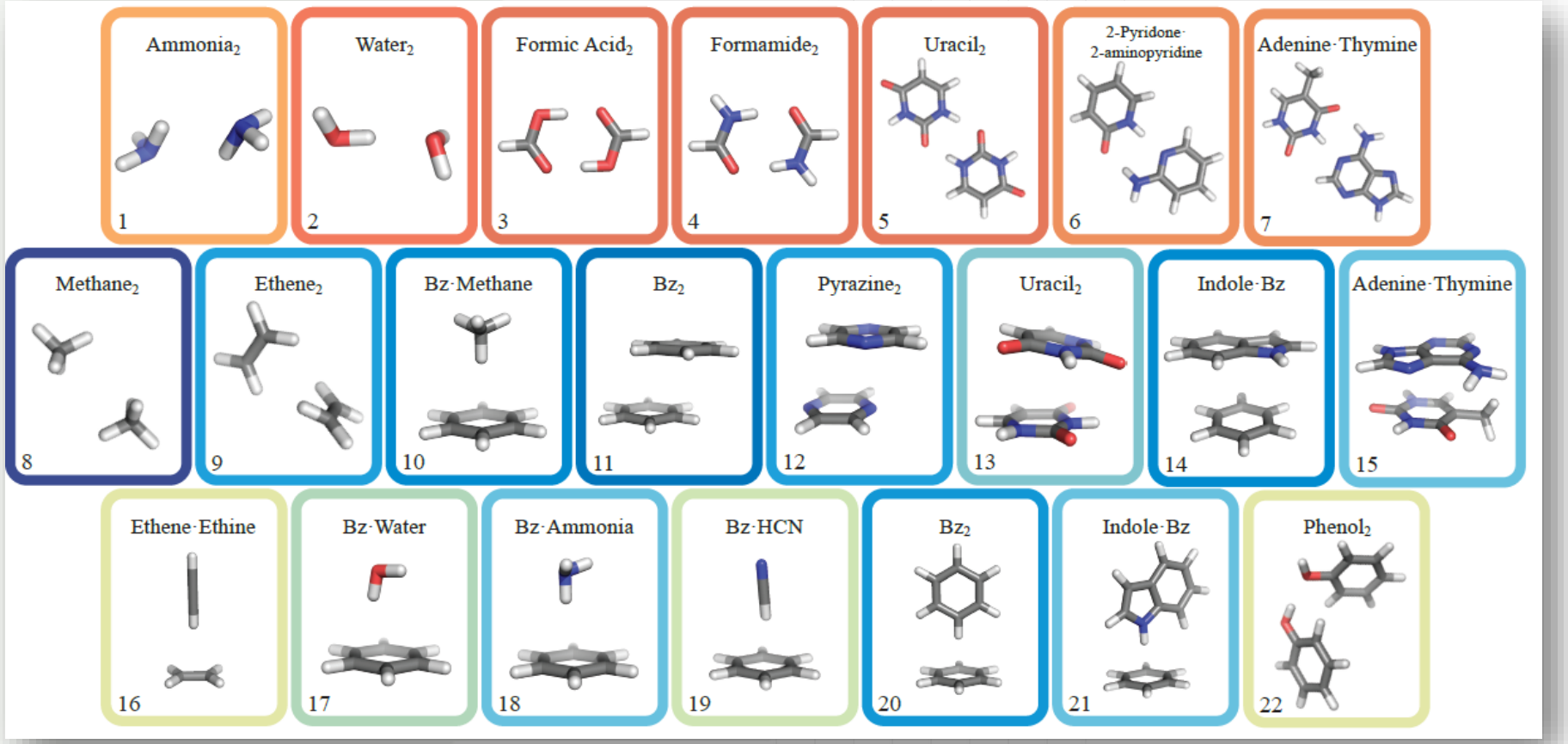

2CJI
-1.30 kcal/mol (-1.72 expt)
PEP216
PEP215
PEP219
PEP214
Ser195
PEP190
PEP227
Asp189
PEP213
PEP226
Tyr228
2PR3
-1.28 (-1.41 expt)
$\Delta\Delta E^{A}_{\mathrm{int}}$
-2.5
kcal mol^{-1}
2.5
2W26
-0.71 (-1.06 expt)
3ENS
-2.80 (-2.31 expt)

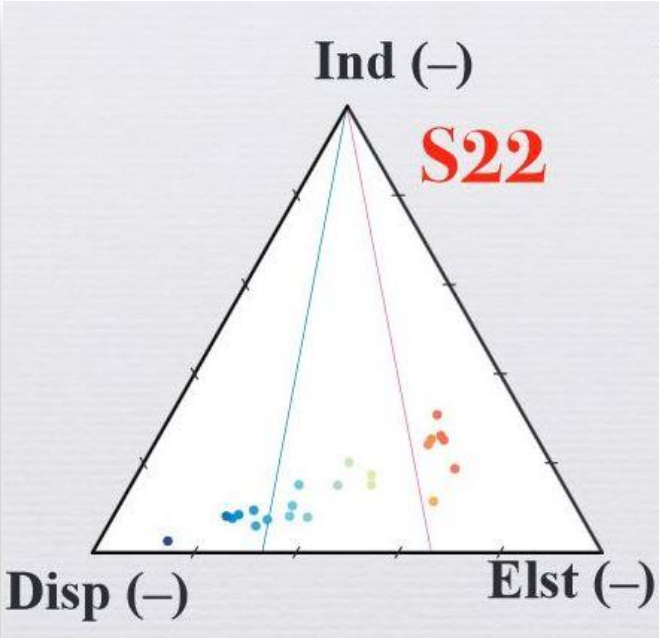
OPPORTUNITIES AND CHALLENGES

- Quantum chemistry can explain why one drug worked better than the other!
- But… the computations took days…
- Goal: develop a much faster (yet accurate!) computational model
- We don’t know what the model should look like… try a data-driven approach based on high-quality quantum chemistry

INTERACTION TYPES

S22 Test Set



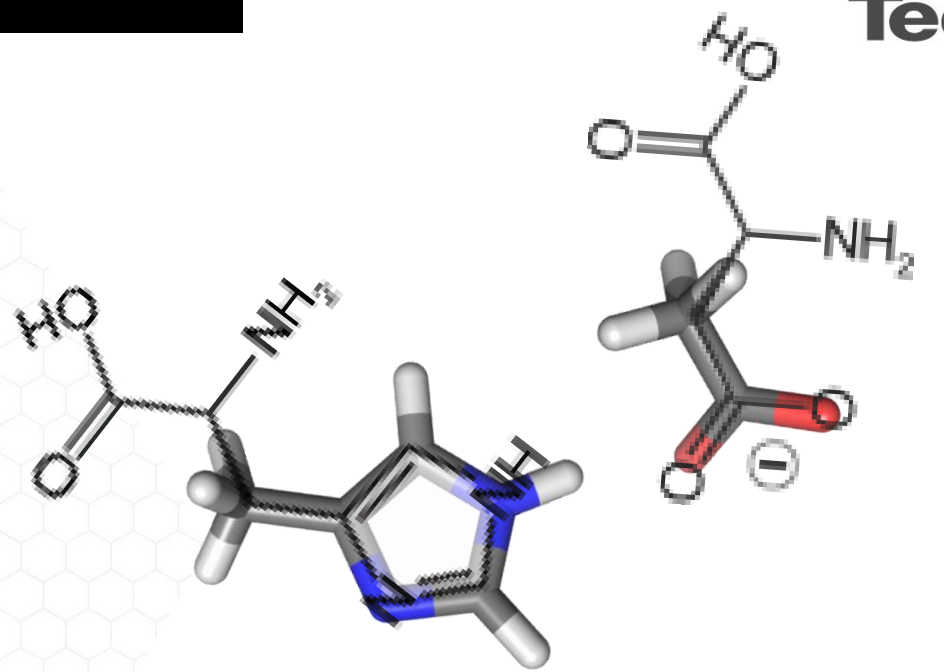


Color coding from symmetry-adapted perturbation theory (SAPT)

P. Jurecka, J. Sponer, J. Cerny, and P. Hobza, *Phys. Chem. Chem. Phys.* **8**, 1985 (2006)

THE BIO-FRAGMENT DATABASE (BFDB)

- Benchmark-quality reference values for nearly all types of non-covalent contacts found in the Protein Data Bank (PDB)
- Nearly-redundant contacts filtered out



Datasets:

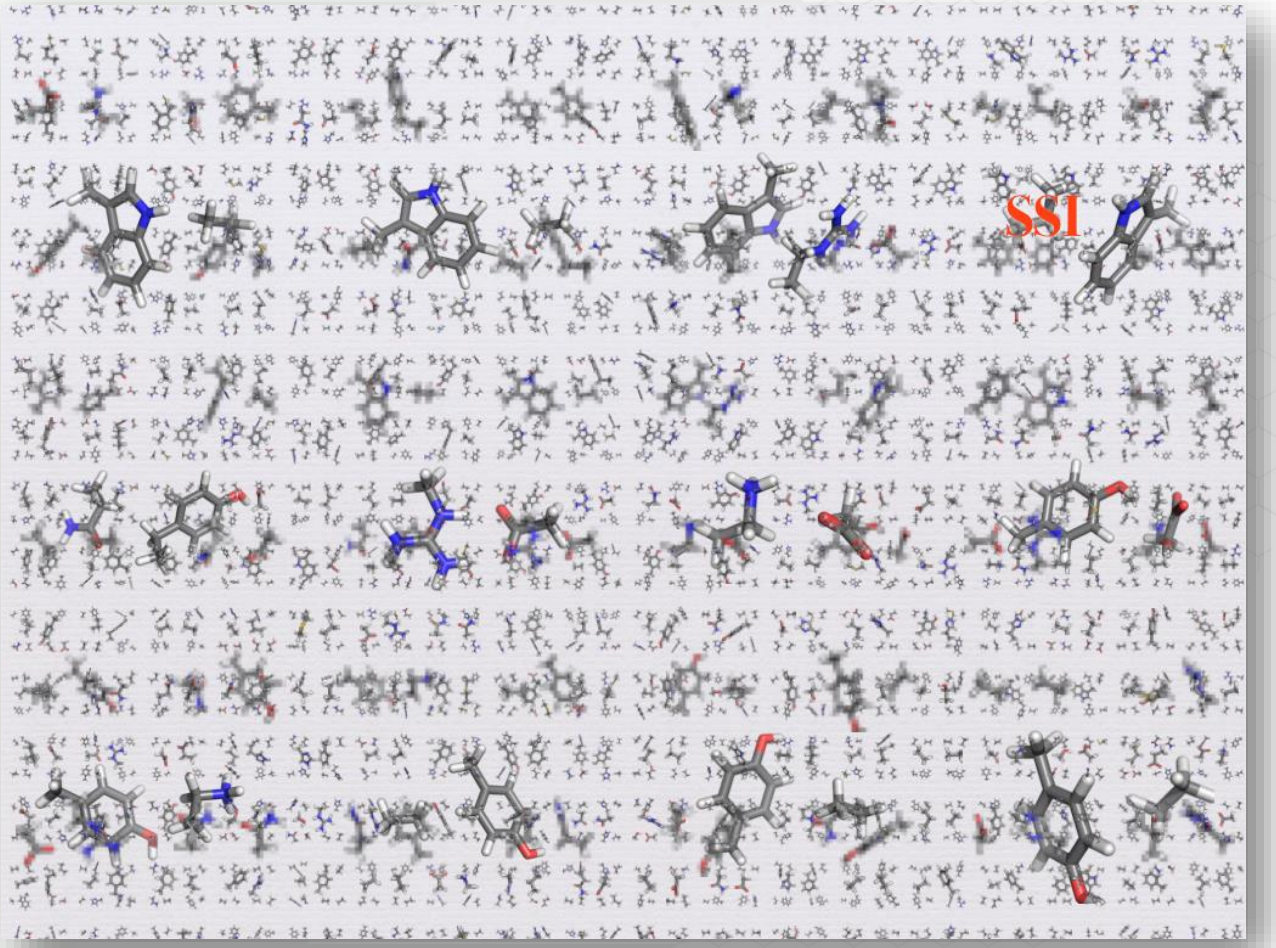
- Side-chain/side-chain (SSI): 3384
- Backbone-backbone (BBI): 100
- Sidechain-backbone (future, SBI): 2774

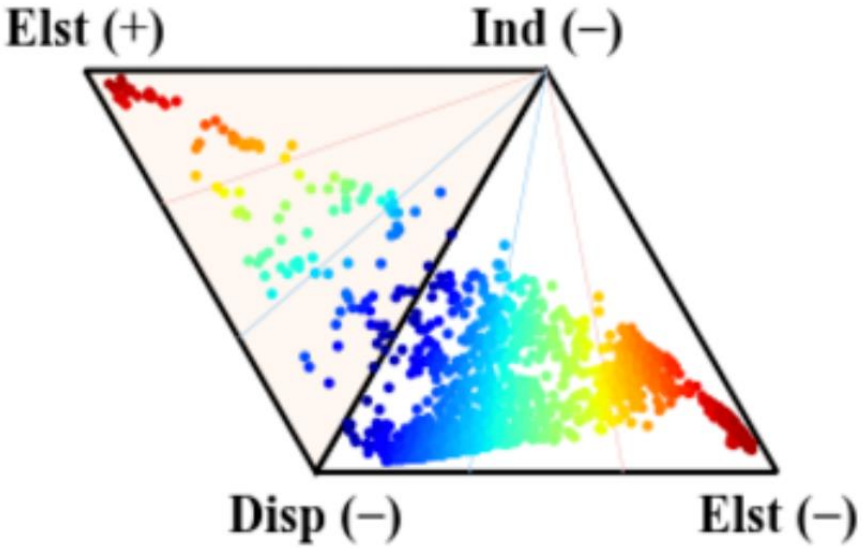
Collaboration with Kennie Merz & Alex MacKerell

L. A. Burns et al., *J. Chem. Phys*. **147**, 161727 (2017)

INTERACTION TYPES

SSI Test Set

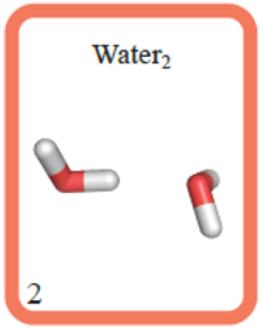




THE STUDY

- Tested various approximate solutions to see how well they compared to the highly-accurate reference data
- 384 approximate methods were examined for part or all of the 3484 dimers in BBI and SSI
- Total of >1M data points

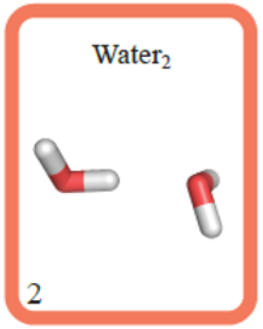
OLD WORKFLOW




```
$molecule
0 1
O -1.485 -0.115  0.000
H -1.868  0.762  0.000
H -0.534  0.041  0.000
O  1.416  0.111  0.000
H  1.746 -0.374 -0.759
H  1.746 -0.374  0.759
$end

$rem
BASIS        cc-pVDZ
EXCHANGE HF
JOBTYPE      SP
$end
```

Input File
(Q-Chem program)

OLD WORKFLOW



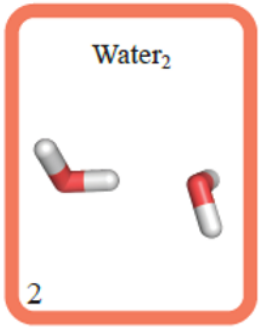

```
$molecule
0 1
O -1.485 -0.115  0.000
H -1.868  0.762  0.000
H -0.534  0.041  0.000
O  1.416  0.111  0.000
H  1.746 -0.374 -0.759
H  1.746 -0.374  0.759
$end

$rem
BASIS        cc-pVDZ
EXCHANGE HF
JOBTYPE      SP
$end
```

Input File
(Q-Chem program)

```
....
Hartree-Fock
A restricted SCF calculation will be
performed using DIIS
SCF converges when DIIS error is below 1.0e-05
----------------------------------------
Cycle      Energy        DIIS error
----------------------------------------
  1   -151.8599071182     9.96e-02
  2   -152.0015355569     1.61e-02
  3   -152.0498457519     7.88e-03
  4   -152.0621557925     9.77e-04
  5   -152.0625062601     2.39e-04
  6   -152.0625348103     4.92e-05
  7   -152.0625361356     1.80e-05
  8   -152.0625362323     3.87e-06  Convergence criterion met
----------------------------------------
SCF time:   CPU 0.34s  wall 1.00s
SCF   energy in the final basis set =    -152.0625362323
Total energy in the final basis set =   -152.0625362323
....
```

Output File

OLD WORKFLOW




```
$molecule
0 1
O -1.485 -0.115  0.000
H -1.868  0.762  0.000
H -0.534  0.041  0.000
O  1.416  0.111  0.000
H  1.746 -0.374 -0.759
H  1.746 -0.374  0.759
$end

$rem
BASIS        cc-pVDZ
EXCHANGE HF
JOBTYPE      SP
$end
```

Input File
(Q-Chem program)

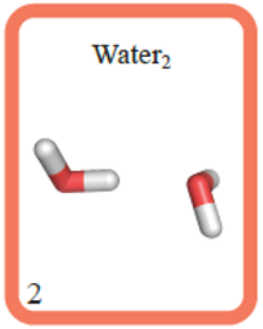
```
....
Hartree-Fock
A restricted SCF calculation will be
performed using DIIS
SCF converges when DIIS error is below 1.0e-05
---------------------------------------
Cycle      Energy        DIIS error
---------------------------------------
  1   -151.8599071182     9.96e-02
  2   -152.0015355569     1.61e-02
  3   -152.0498457519     7.88e-03
  4   -152.0621557925     9.77e-04
  5   -152.0625062601     2.39e-04
  6   -152.0625348103     4.92e-05
  7   -152.0625361356     1.80e-05
  8   -152.0625362323     3.87e-06  Convergence criterion met
---------------------------------------
SCF time:   CPU 0.34s  wall 1.00s
SCF   energy in the final basis set =    -152.0625362323
Total energy in the final basis set =    -152.0625362323
....
```

Output File

	A	B	C	D	E	F
1	H2O Dimer interaction energy w/ CP-correction					
2						
3	Dimer	MonoA	MonoB	IE (E_h)	IE (kcal/mol)	
4	-156.062536					
5						
6						

Spreadsheet

OLD WORKFLOW




```
$molecule
0 1
O -1.485 -0.115  0.000
H -1.868  0.762  0.000
H -0.534  0.041  0.000
@O  1.416  0.111  0.000
@H  1.746 -0.374 -0.759
@H  1.746 -0.374  0.759
$end

$rem
BASIS        cc-pVDZ
EXCHANGE HF
JOBTYPE      SP
$end
```

Input File
(Q-Chem program)

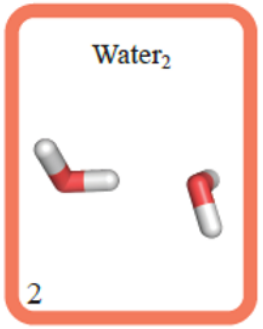
```
....
Hartree-Fock
A restricted SCF calculation will be
performed using DIIS
SCF converges when DIIS error is below 1.0e-05
---------------------------------------
Cycle      Energy        DIIS error
---------------------------------------
  1    -75.9081353824     6.97e-02
  2    -75.9930406969     1.20e-02
  3    -76.0195407670     6.06e-03
  4    -76.0267627057     6.91e-04
  5    -76.0269358630     1.80e-04
  6    -76.0269509986     3.02e-05
  7    -76.0269515356     4.63e-06  Convergence criterion met
---------------------------------------
SCF time:   CPU 0.22s  wall 0.00s
SCF   energy in the final basis set =      -76.0269515356
Total energy in the final basis set =      -76.0269515356
....
```

Output File

	A	B	C	D	E	F
1	H2O Dimer interaction energy w/ CP-correction					
2						
3	Dimer	MonoA	MonoB	IE (E_h)	IE (kcal/mol)	
4	-156.062536	-76.026952				
5						
6						

Spreadsheet

OLD WORKFLOW




```
$molecule
0 1
O -1.485 -0.115  0.000
H -1.868  0.762  0.000
H -0.534  0.041  0.000
@O  1.416  0.111  0.000
@H  1.746 -0.374 -0.759
@H  1.746 -0.374  0.759
$end

$rem
BASIS        cc-pVDZ
EXCHANGE HF
JOBTYPE      SP
$end
```

Input File
(Q-Chem program)

```
....
Hartree-Fock
A restricted SCF calculation will be
performed using DIIS
SCF converges when DIIS error is below 1.0e-05
---------------------------------------
Cycle      Energy        DIIS error
---------------------------------------
  1    -75.9109558111     6.99e-02
  2    -75.9961062627     1.19e-02
  3    -76.0221162806     6.15e-03
  4    -76.0295303742     6.94e-04
  5    -76.0297018170     1.79e-04
  6    -76.0297161408     2.87e-05
  7    -76.0297166309     4.91e-06  Convergence criterion met
---------------------------------------
SCF time:   CPU 0.23s  wall 0.00s
SCF   energy in the final basis set =     -76.0297166309
Total energy in the final basis set =     -76.0297166309
```

Output File

	A	B	C	D	E	F
1	H2O Dimer interaction energy w/ CP-correction					
2						
3	Dimer	MonoA	MonoB	IE (E_h)	IE (kcal/mol)	
4	-152.062536	-76.026952	-76.029717	-0.005867	-3.68	
5						

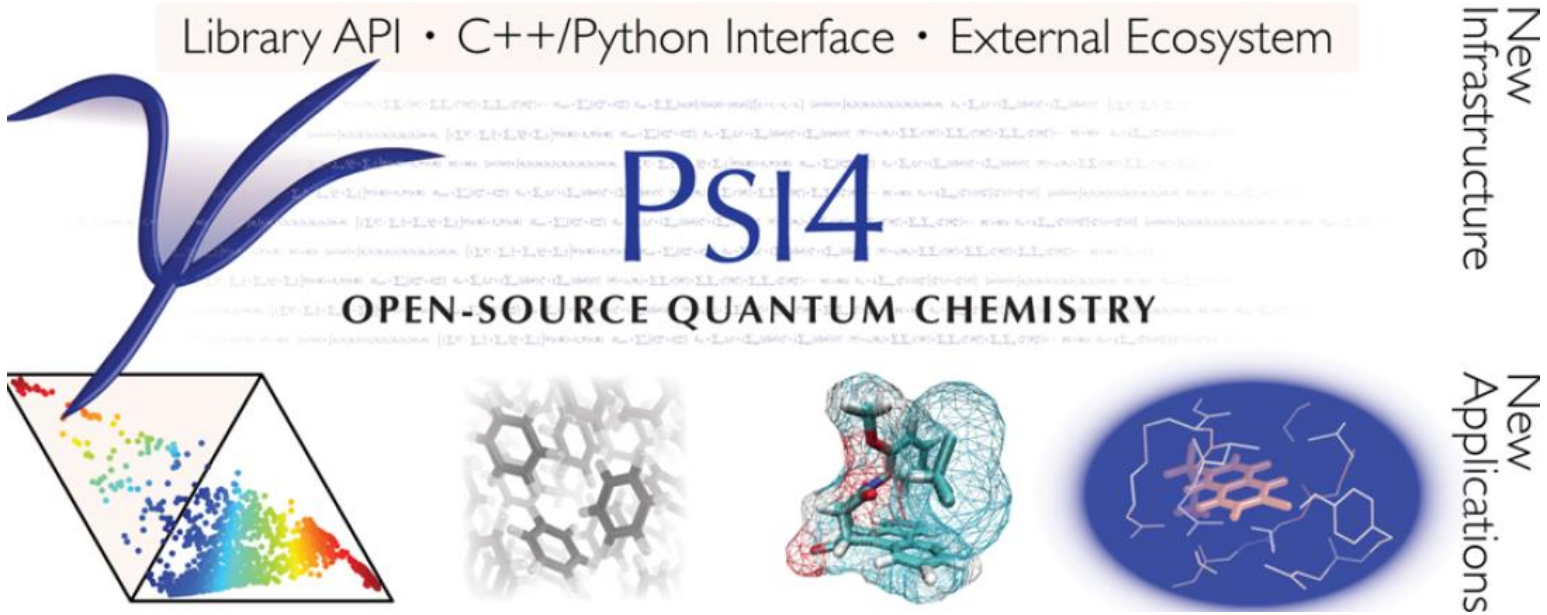
Spreadsheet

Final
Result

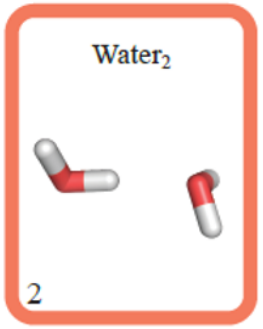
JCTC Journal of Chemical Theory and Computation
Article
pubs.acs.org/JCTC

Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability

Robert M. Parrish,[†] Lori A. Burns,[†] Daniel G. A. Smith,[†] Andrew C. Simmonett,[‡] A. Eugene DePrince, III,[¶] Edward G. Hohenstein,[§] Uğur Bozkaya,[∥] Alexander Yu. Sokolov,[⊥] Roberto Di Remigio,[#] Ryan M. Richard,[†] Jérôme F. Gonthier,[†] Andrew M. James,[@] Harley R. McAlexander,[@] Ashutosh Kumar,[@] Masaaki Saitow,[△] Xiao Wang,[@] Benjamin P. Pritchard,[†] Prakash Verma,[▽] Henry F. Schaefer, III,[○] Konrad Patkowski,[◆] Rollin A. King,[&] Edward F. Valeev,[@] Francesco A. Evangelista,[▽] Justin M. Turney,[○] T. Daniel Crawford,[@] and C. David Sherrill[*,†]



NEW WORKFLOW WITH PSI4




```
molecule {
0 1
O -1.485 -0.115  0.000
H -1.868  0.762  0.000
H -0.534  0.041  0.000
--
O  1.416  0.111  0.000
H  1.746 -0.374 -0.759
H  1.746 -0.374  0.759
}

energy('scf/cc-pVDZ', bsse_type = 'cp')
```

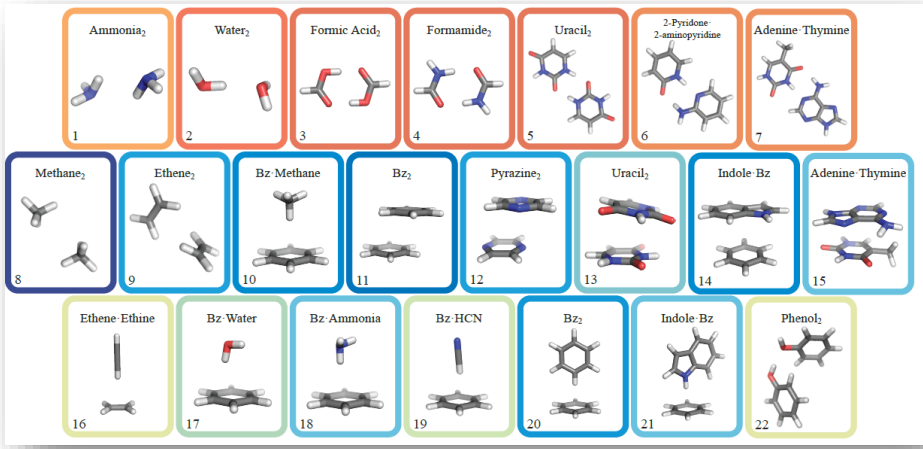
Input File
(Psi4 program)

```
...
==> N-Body: Counterpoise Corrected (CP) energies <==

  n-Body    Total Energy [Eh]      I.E. [kcal/mol]      Delta [kcal/mol]
     1    -152.053271572521      0.000000000000       0.000000000000
     2    -152.059137711711     -3.681057916101      -3.681057916101
...
```

Output File

We automated the counterpoise procedure with a simple Python function…
Now just one computation, not 3! No more spreadsheet!

ENTIRE DATABASES WITH PSI4



Now let's automate running through all dimers in an entire database!

Input File
(Psi4 program)

```
database('scf/cc-pVDZ', 'S22', bsse_type = 'cp')
```

The "database" function is just a Python function!

Output File

==> Requested Energy <==

Reaction	Reaction Energy Ref	Reaction Energy Calc	Reaction Error [kcal/mol]	Reaction Error [kJ/mol]	Reagent 1 [Eh]	Wt	Reagent 2 [Eh]	Wt	Reagent 3 [Eh]	Wt
S22-1	-3.1330	-1.2160	1.9170	8.0206	-112.39621715	1	-56.19713965	-1	-56.19713965	-1
S22-2	-4.9890	-3.6811	1.3079	5.4724	-152.06249065	1	-76.02693046	-1	-76.02969405	-1
S22-3	-18.7530	-14.6225	4.1305	17.2820	-377.58563042	1	-188.78116399	-1	-188.78116399	-1
S22-4	-16.0620	-11.5979	4.4641	18.6777	-337.91567740	1	-168.94859746	-1	-168.94859746	-1

Minimal Dev	0.8972	3.7541
Maximal Dev	15.2329	63.7344
Mean Signed Dev	5.1089	21.3757
Mean Absolute Dev	5.1089	21.3757
RMS Dev	6.2829	26.2877

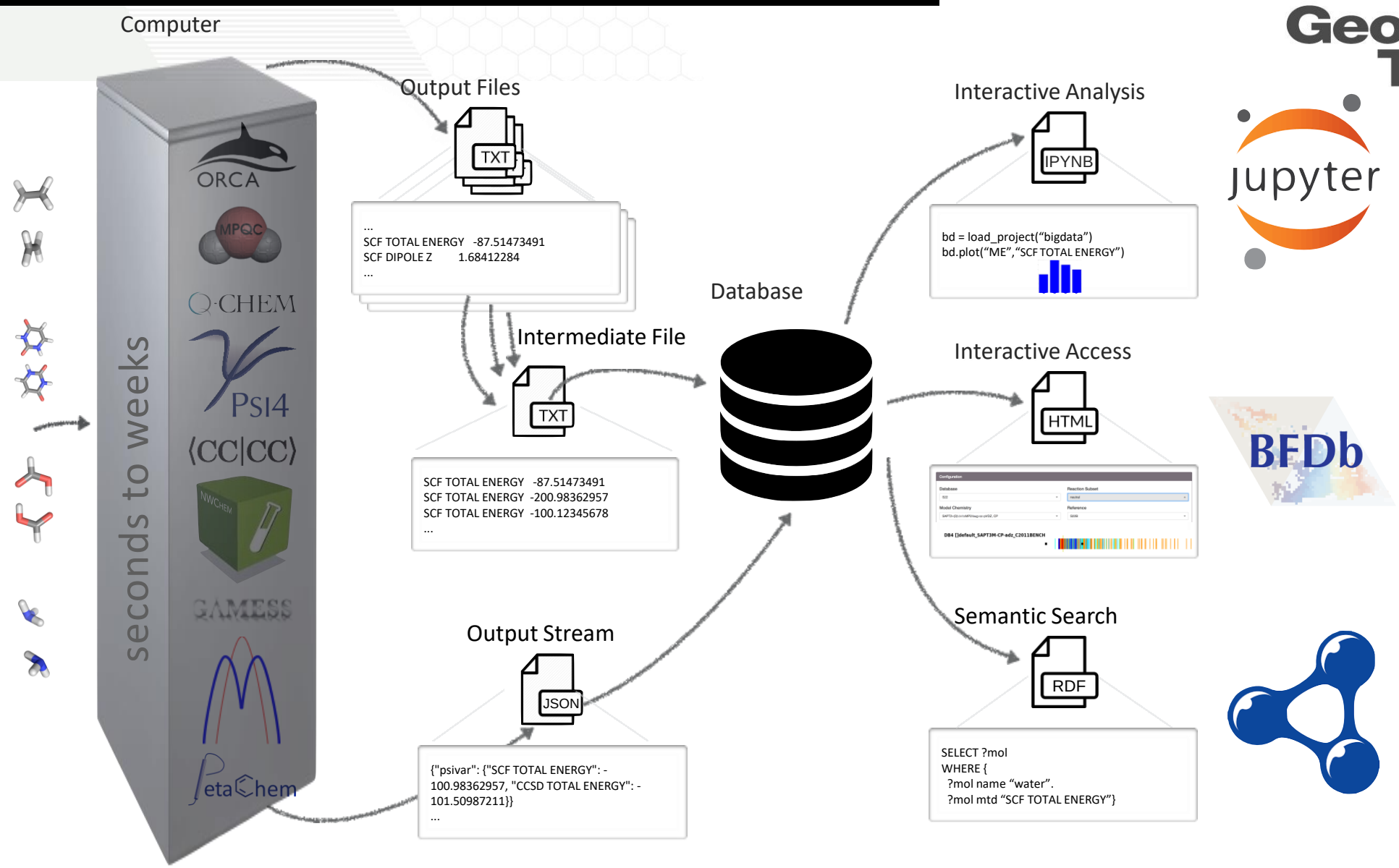
PYTHON ACCESSIBILITY OF DATA

```
def table_4col_spec(**kwargs):
    rowplan = ['bas', 'mtd']
    columnplan = [
        ['l', r"""Method \& Basis Set""", '',                textables.label,   {}],
        ['d', 'Absolute Error',           'total',           textables.val,     {'sset': 'default'}],
        ['d', 'Absolute Error',           'polar',           textables.val,     {'sset': 'polarpolar'}],
        ['d', 'Absolute Error',           r'$\bm{\pi/\pi}$', textables.val,     {'sset': 'arylaryl'}],
        ['d', 'Relative Error',           'total',           textables.val,     {'sset': 'default', 'err': 'mape'}],
        ['d', r"""Iowa""",                '',                textables.liliowa, {}],
        ['d', r"""Error Distribution""",  '',                textables.flat,    {'sset': 'default'}]
    ]

    landscape = False
    footnotes = []
    theme = ''
    title = r"""Interact
    return rowplan, colu

def make_Table_1(dbobj):
    dbobj.table_wrapper(

make_Table_1(ssi)
!pdflatex test1 && open
```

TABLE I. Interaction energy demo

Method & Basis Set	Absolute Error			Relative Error	Iowa	Error Distribution
	total	polar	π/π	total		
aug-cc-pVDZ						
MP2	0.38	0.65	0.40	49.0		
SCS-MP2	0.96	1.28	0.73	119.7		
SCS(MI)-MP2						
aug-cc-pVTZ						
MP2	0.24	0.26	0.62	21.9		
SCS-MP2	0.71	0.93	0.51	87.9		
SCS(MI)-MP2	0.22	0.22	0.37	35.4		
aug-cc-pVQZ						
MP2	0.22	0.13	0.69	17.2		
SCS-MP2	0.63	0.79	0.43	80.0		
SCS(MI)-MP2	0.29	0.30	0.17	50.9		

DATA FLOW IN MODERN QUANTUM CHEMISTRY



Georgia Tech

jupyter

BFDb

ACKNOWLEDGMENTS