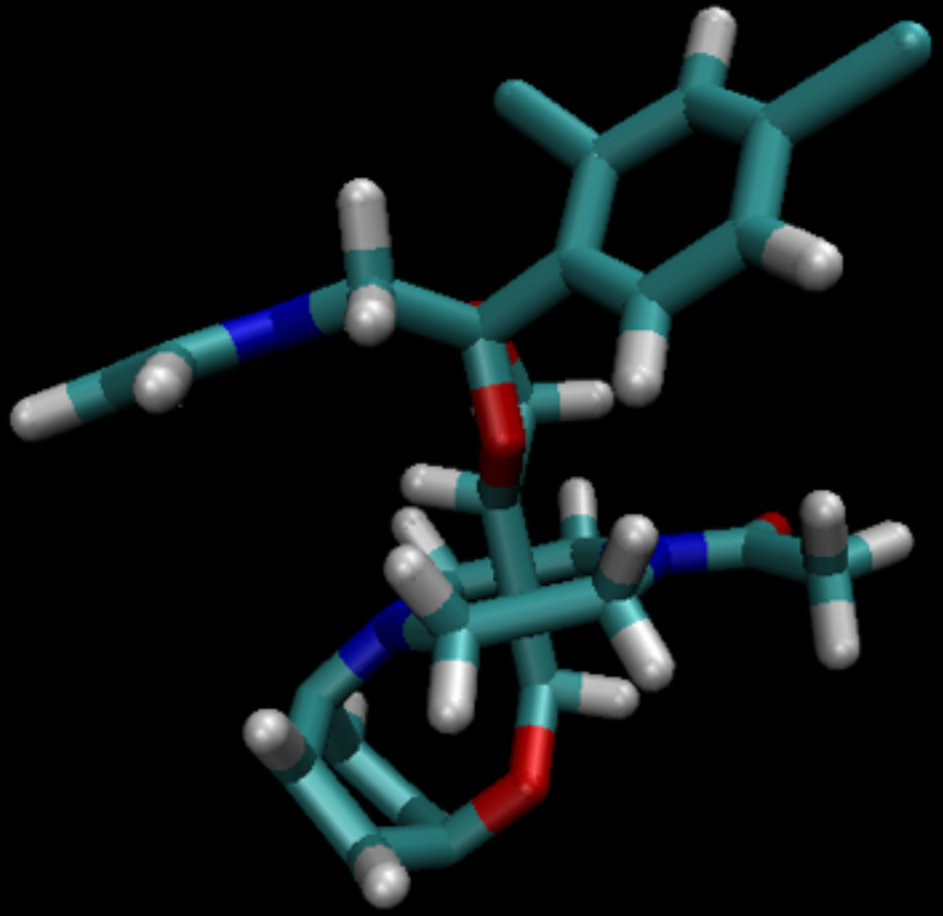
Difficult lessons learned from QM & pK_a calculations in SAMPL5

Frank C. Pickard IV

D3R Meeting

11 March 2016



A “high energy” conformation of molecule 92.

SAMPL5 Description

A good proxy measurement for drug availability.

This is a "broad" definition of cosolvation, it includes ionic species, microsolvation, tautomers, snorkeling, *etc.*

$$\begin{aligned}\log D_{\text{chex/aq}}(x) &= \log_{10}\left(\frac{[x]_{\text{chex}}}{[x]_{\text{aq}} + [x']_{\text{aq}} + [x'']_{\text{aq}} + \ldots}\right)\\ &= (\Delta G_{\text{aq}} - \Delta G_{\text{chex}})\frac{\log_{10}(e)}{kT}\end{aligned}$$

1 log D = 1.36 kcal/mol

SAMPL5 Methods: log P calculation

- MM BAR
- QM Optimization
- QM NBB
- QM/MM Zwanzig
- Semi-Empirical NBB (in progress)

$$\log P_{\mathrm{chex/aq}}(x) = (\Delta G_{\mathrm{aq}} - \Delta G_{\mathrm{chex}}) \frac{\log_{10}(e)}{kT}$$

SAMPL5 Methods: log P calculation

Non-Boltzmann Bennett Method

$$\Delta A = k_\mathrm{B} T \ln\left(\frac{\langle f(U_0 - U_1 + C)\rangle_1}{\langle f(U_1 - U_0 - C)\rangle_0}\right) + C \xrightarrow{V^b}$$

$$\langle X\rangle_\mathrm{unbiased} = \frac{\langle X \exp(V^b/k_\mathrm{B}T)\rangle_\mathrm{biased}}{\langle \exp(V^b/k_\mathrm{B}T)\rangle_\mathrm{biased}}$$

$$\Delta A = k_\mathrm{B} T \ln\left(\frac{\langle f(U_0 - U_1 + C)\exp(V_1^b/k_\mathrm{B}T)\rangle_1 \langle \exp(V_0^b/k_\mathrm{B}T)\rangle_0}{\langle f(U_1 - U_0 - C)\exp(V_0^b/k_\mathrm{B}T)\rangle_0 \langle \exp(V_1^b/k_\mathrm{B}T)\rangle_1}\right) + C$$

$$f(x) = \frac{1}{1 + \exp(x/k_\mathrm{B}T)}$$

König and Boresch J. Comp. Chem. (2010) 32, 1082.

SAMPL5 Methods: log P calculation

QM/MM Non-Boltzmann Bennett

$$V^b = U_{\mathrm{MM}} - U_{\mathrm{QM}}$$

$$\Delta A = k_{\mathrm{B}}T \ln\left(\frac{\langle f(U_{0,\mathrm{QM}} - U_{1,\mathrm{QM}} + C)\rangle_{1,\mathrm{MM}}}{\langle f(U_{1,\mathrm{QM}} - U_{0,\mathrm{QM}} - C)\rangle_{0,\mathrm{MM}}}\right) + C$$

$$\Delta A = k_{\mathrm{B}}T \ln\left(\frac{\langle f(U_{0,\mathrm{QM}} - U_{1,\mathrm{QM}} + C)\exp(V_1^b/k_{\mathrm{B}}T)\rangle_{1,\mathrm{MM}}\langle \exp(V_0^b/k_{\mathrm{B}}T)\rangle_{0,\mathrm{MM}}}{\langle f(U_{1,\mathrm{QM}} - U_{0,\mathrm{QM}} - C)\exp(V_0^b/k_{\mathrm{B}}T)\rangle_{0,\mathrm{MM}}\langle \exp(V_1^b/k_{\mathrm{B}}T)\rangle_{1,\mathrm{MM}}}\right) + C$$

König, Pickard, Mei and Brooks J. Comput. Aided Mol. Des. (2014) 28, 245.

SAMPL5 Methods: log P calculation

- CGenFF
- HREX Simulations, LD NVT
- 36 lambda points (6 electrostatic, 30 vdw)
- 1 fs timestep, 5 ns total
- 5000 QM or QM/MM calculations

SAMPL5 Methods: log P calculation

- QM Optimization (our "control" submission)
    - w/ SMD Implicit Solvent (Vertical or Relaxed Solvation)
    - M06-2X/6-311++G**/6-31+G* with SMD
- QM NBB (optimized from SAMPL4 data)
    - w/ SMD Implicit Solvent
    - M06-2X/6-31+G* **or** OLYP/DZP
- QM/MM Zwanzig
    - w/ TIP3P Explicit Solvent
    - BLYP/6-31G*

SAMPL5 Methods: log D correction

- p*K*a corrections
  - absolute/relative
  - vertical/relaxed solvation
- tautomerization (we only looked at aqueous)
- dimerization (in progress), trimerization, *etc.*
- wet cyclohexane

$$\log D_{\mathrm{chex/aq}}(x) = \log P_{\mathrm{chex/aq}}(x) + \boldsymbol{\Delta}\mathbf{G}_{\mathrm{corr}} \frac{\log_{10}(e)}{kT}$$

$$\Delta G_{\mathrm{corr}} = \Delta G_{\mathrm{p}K_{\mathrm{a}}} + \Delta G_{\mathrm{taut}} + \Delta G_{\mathrm{dimer}} + \Delta G_{\mu-\mathrm{solv}} + \ldots$$

SAMPL5 Methods: p*K*a correction

Deprotonation Thermocycle:

$$\begin{tikzcd}
\mathrm{AH}^+_{(\mathrm{g})} \arrow[r, "\Delta G^\circ_{\mathrm{gas}}(\mathrm{AH}^+)"] \arrow[d, "\Delta G^*_{\mathrm{solv}}(\mathrm{AH}^+)"] & \mathrm{A}_{(\mathrm{g})} \arrow[d, "\Delta G^*_{\mathrm{solv}}(\mathrm{A})"] & + & \mathrm{H}^+_{(\mathrm{g})} \arrow[d, "\Delta G^*_{\mathrm{solv}}(\mathrm{H}^+)"] \\
\mathrm{AH}^+_{(\mathrm{aq})} \arrow[r, "\Delta G^\circ_{\mathrm{aq}}(\mathrm{AH}^+)"] & \mathrm{A}_{(\mathrm{aq})} & + & \mathrm{H}^+_{(\mathrm{aq})}
\end{tikzcd}$$

Convert to p*K*a:

$$\Delta G^\circ_{\mathrm{aq}}(AH^+) = \ln(10)RT\mathrm{p}K_{\mathrm{a}} + \Delta G^{\circ\rightarrow *}$$

Convert to populations at pH = 7.4:

$$\mathrm{pH} = \mathrm{p}K_a + \log_{10}\left(\frac{[\mathrm{A}_{(\mathrm{aq})}]}{[\mathrm{A}^+_{(\mathrm{aq})}]}\right)$$

Convert to free energy of protonation…

SAMPL5 Methods: pK_a correction

$$\begin{tikzcd}
\mathrm{AH}^{+}_{(\mathrm{g})} \arrow[r, "\Delta G^{\circ}_{\mathrm{gas}}(\mathrm{AH}^{+})"] \arrow[d, "\Delta G^{*}_{\mathrm{solv}}(\mathrm{AH}^{+})"] & \mathrm{A}_{(\mathrm{g})} \arrow[d, "\Delta G^{*}_{\mathrm{solv}}(\mathrm{A})"] & + & \mathrm{H}^{+}_{(\mathrm{g})} \arrow[d, "\Delta G^{*}_{\mathrm{solv}}(\mathrm{H}^{+})"] \\
\mathrm{AH}^{+}_{(\mathrm{aq})} \arrow[r, "\Delta G^{\circ}_{\mathrm{aq}}(\mathrm{AH}^{+})"] & \mathrm{A}_{(\mathrm{aq})} & + & \mathrm{H}^{+}_{(\mathrm{aq})}
\end{tikzcd}$$

Problematic Terms

Absolute pK_a calculations:

- Proton solvation free energy from experiment (265.9 kcal/mol)
    - Small experimental errors can yield **big** pK_a errors!
- Robustly treat molecules with coupled protonation/tautomerization
- Shouldn’t use with vertical solvation (expensive)

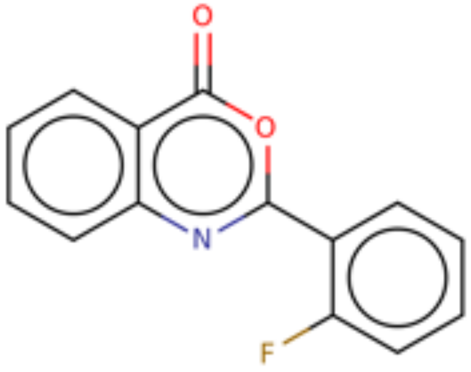
SAMPL5 Methods: p*K*a correction

$$\mathrm{p}K_a = \underbrace{\widetilde{\mathrm{p}K_a}}_{\boxed{\text{Analogue experiment}}} + \left[G(\mathrm{A}_{(\mathrm{aq})}) - G(\mathrm{AH}^+_{(\mathrm{aq})}) - \underbrace{G(\mathrm{L}_{(\mathrm{aq})}) + G(\mathrm{LH}^+_{(\mathrm{aq})})}_{\boxed{\text{Analogue calculation}}}\right] / \left[\ln(10)RT\right]$$

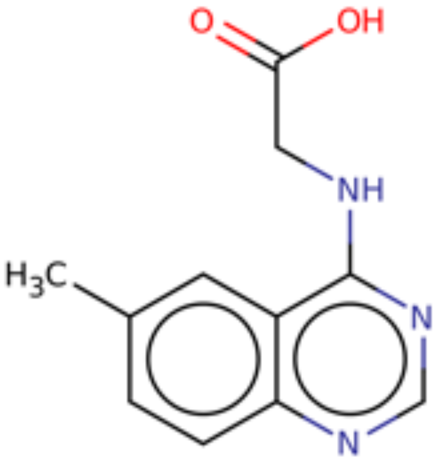
Relative p*K*a calculations:

- Experimental proton solvation free energy term drops out.
    - Uncertainty from analogue experiment
    - Results sensitive to analogue choice
- Can be more accurate than absolute calculations
- Can be used with vertical solvation (cheap)

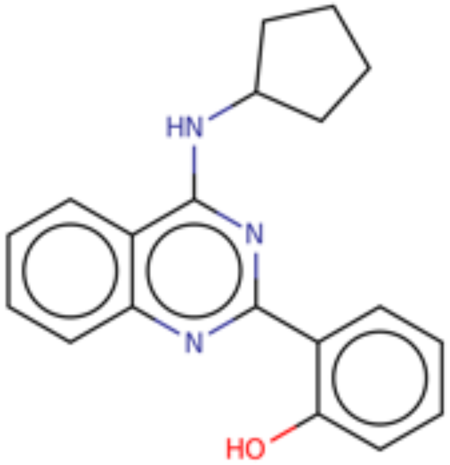
Test Set: Cohort0



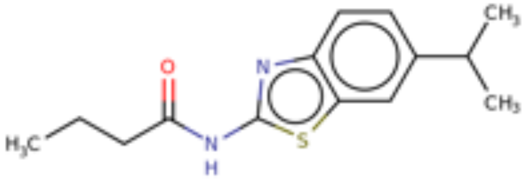
03



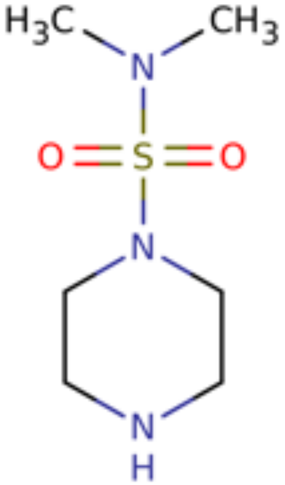
15



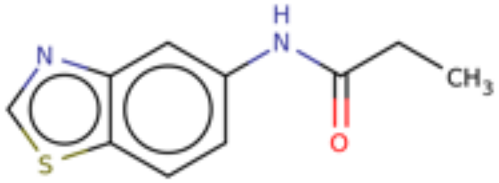
17



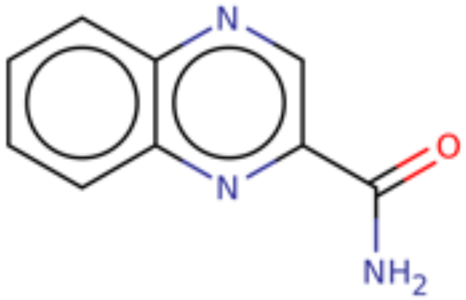
20



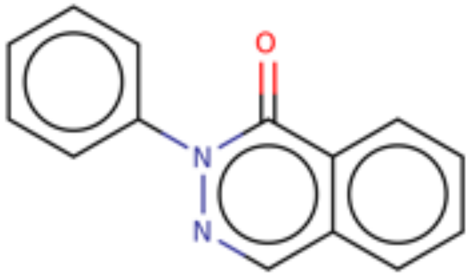
37



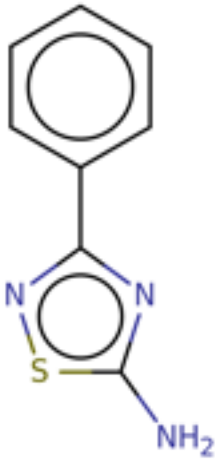
45



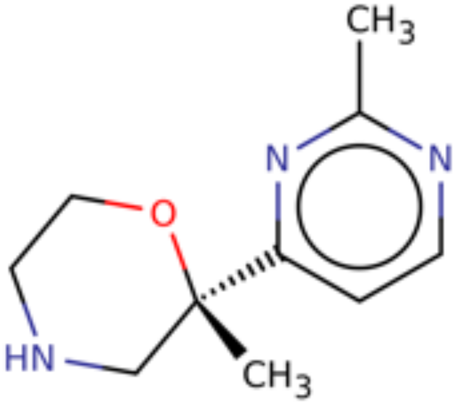
55



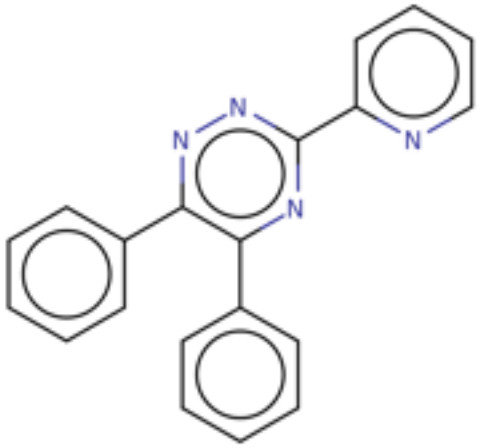
58



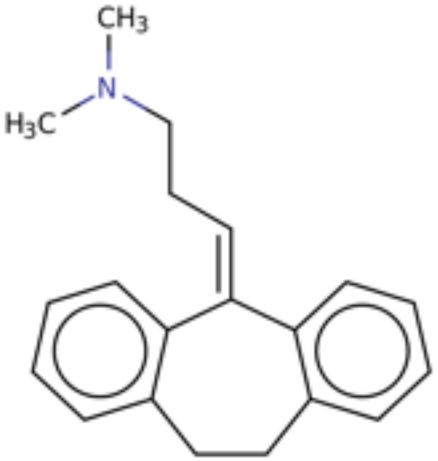
59



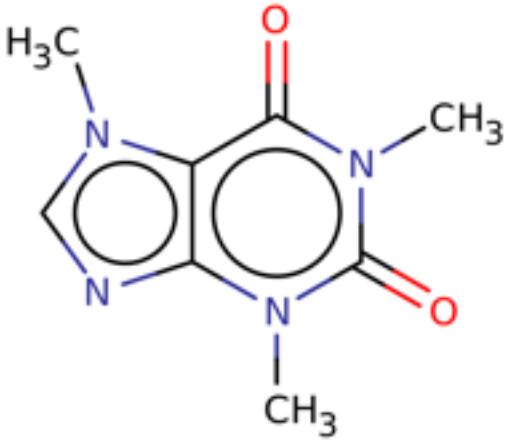
61



68

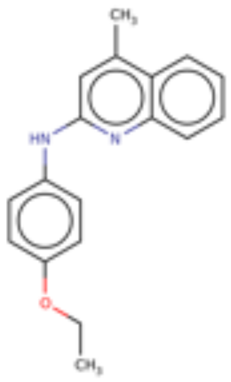


70

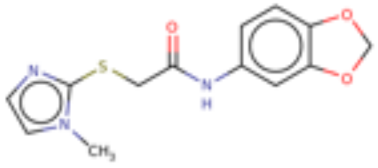


80

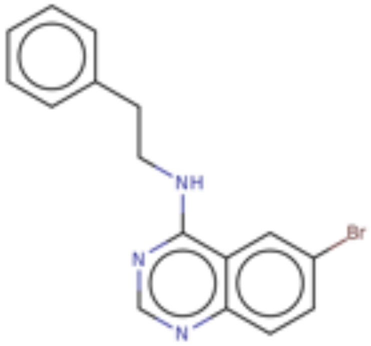
Test Set: Cohort1



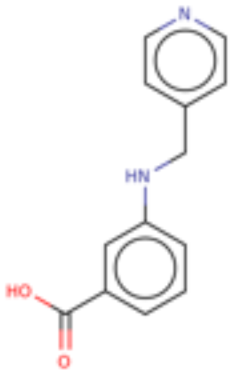
04



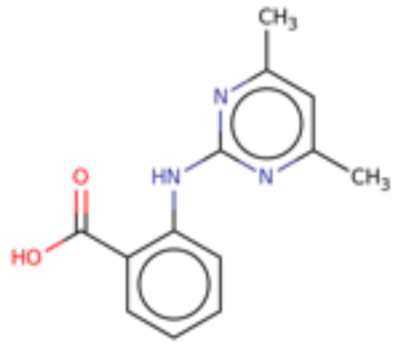
05



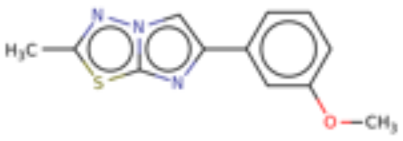
07



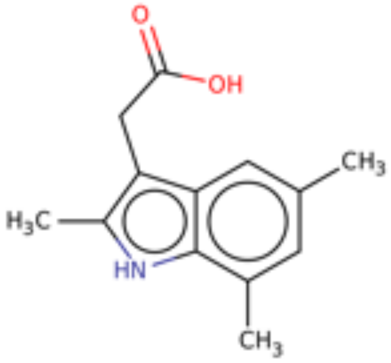
10



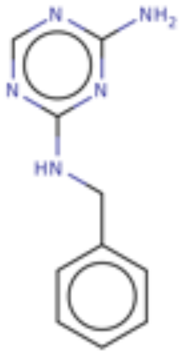
11



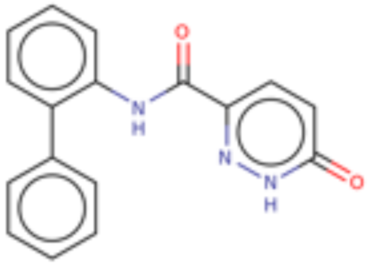
21



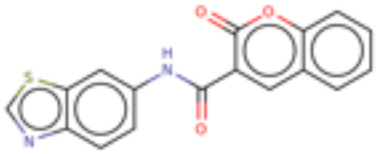
26



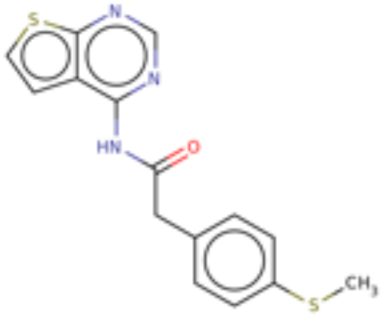
27



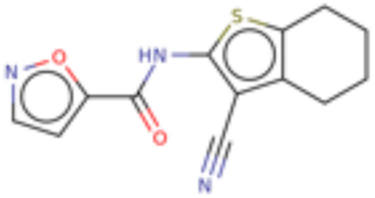
42



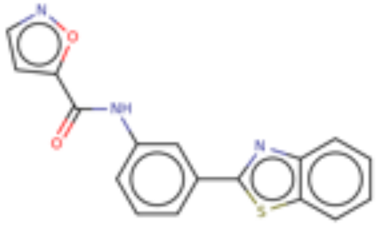
44



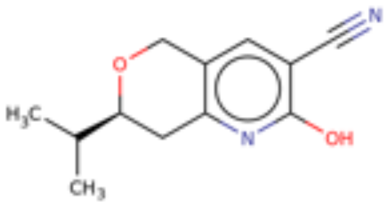
46



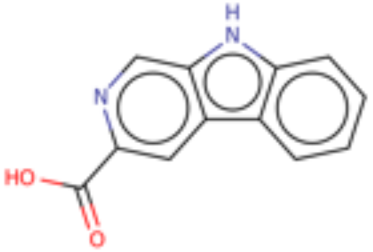
47



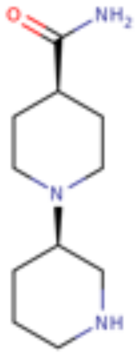
48



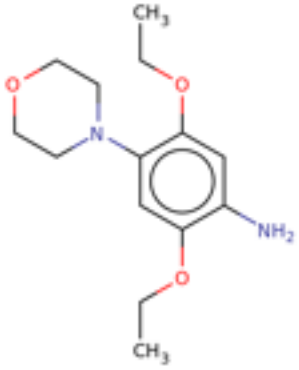
56



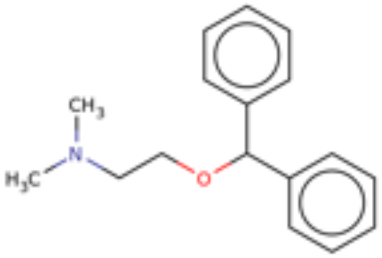
60



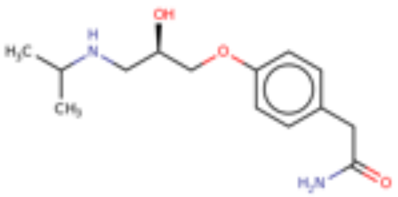
63



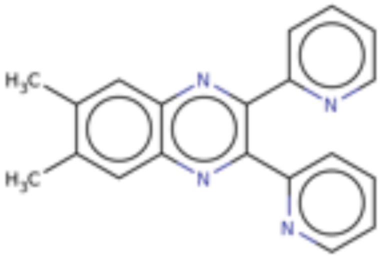
71



72

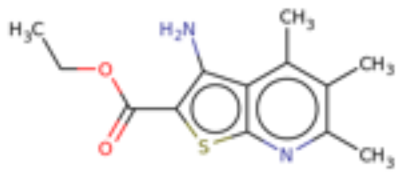


81

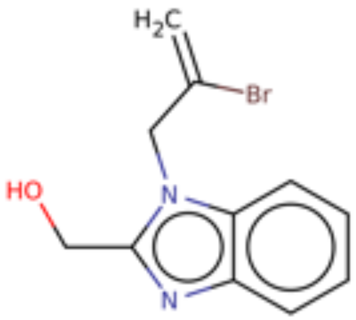


90

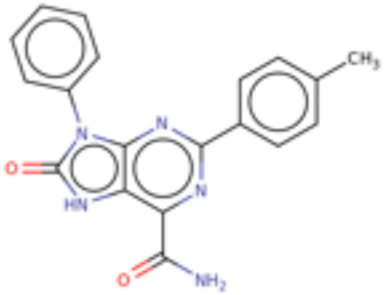
Test Set: Cohort2



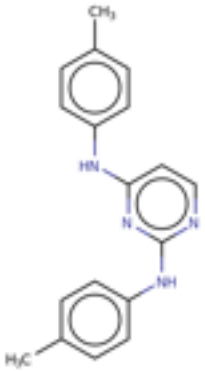
02



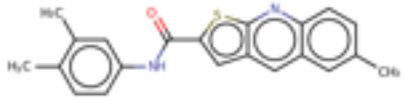
06



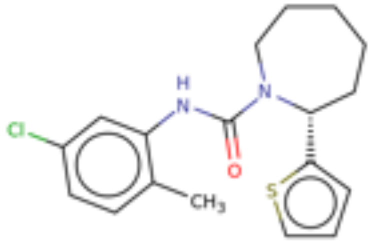
13



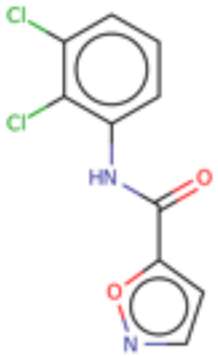
19



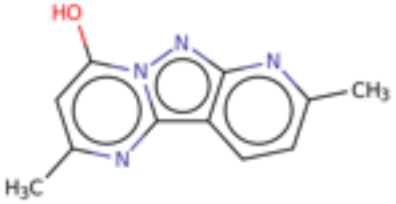
24



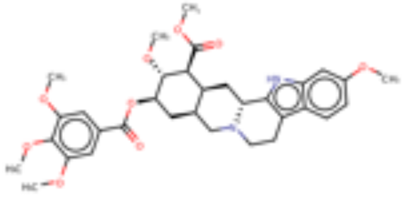
33



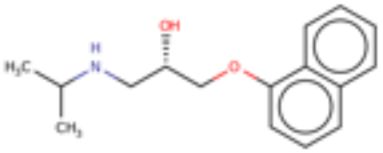
49



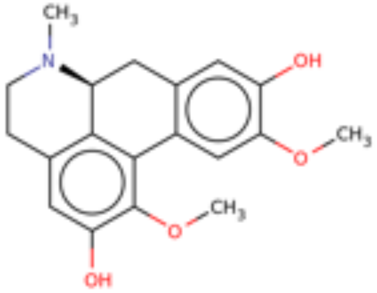
50



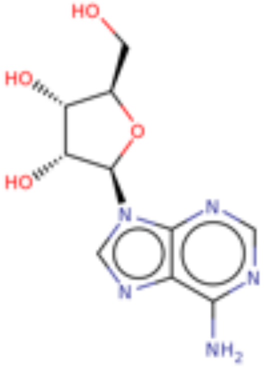
65



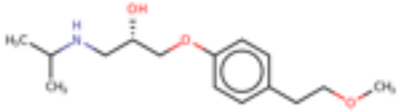
67



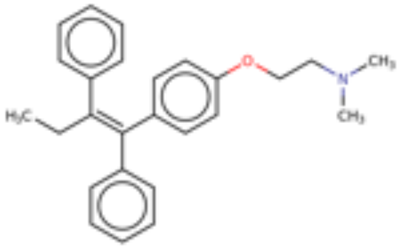
69



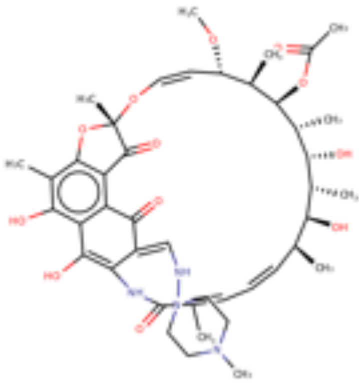
74



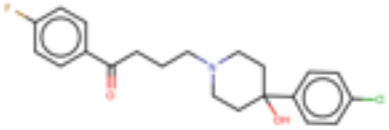
75



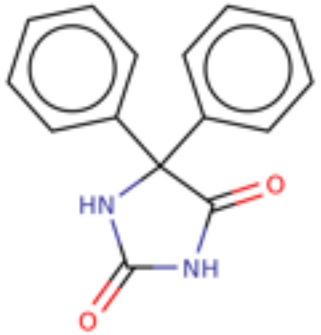
82



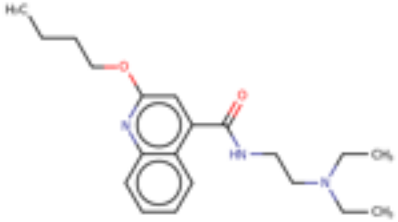
83



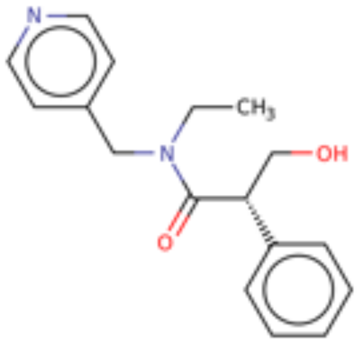
84



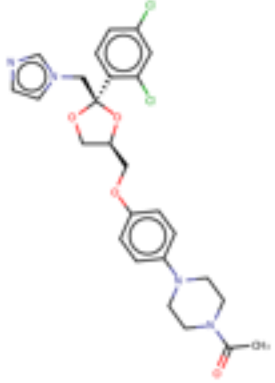
85



86

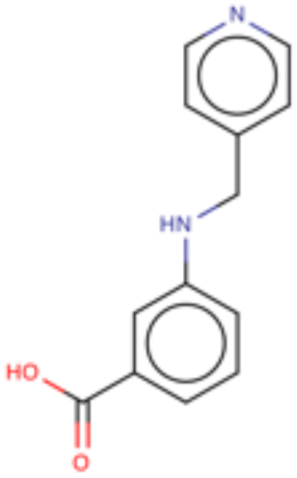


88

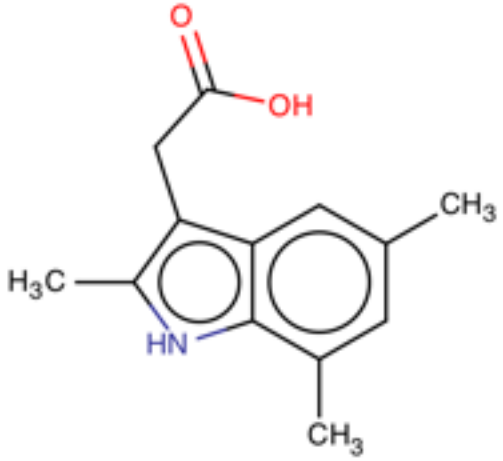


92

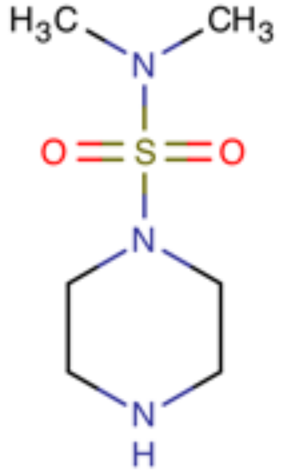
Test Set: pK_a Baddies (simple)



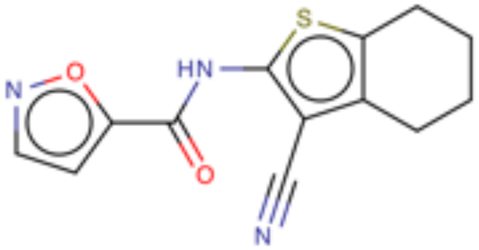
10



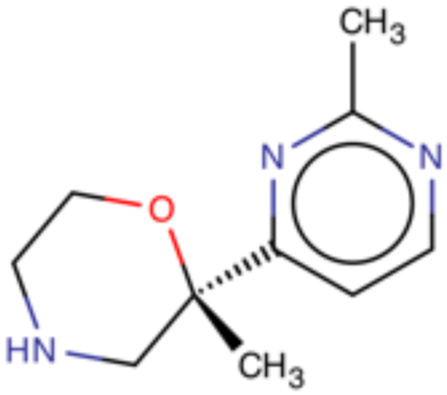
26



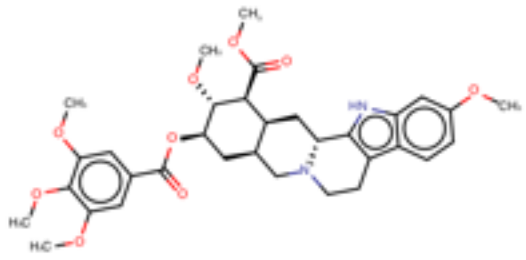
37



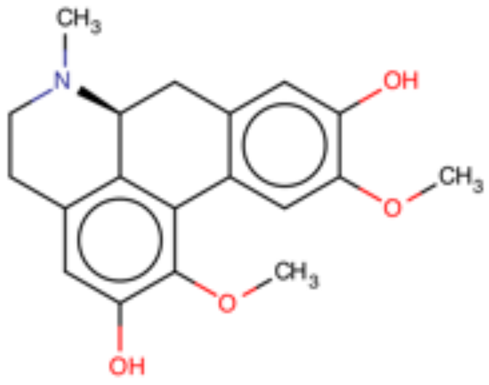
47



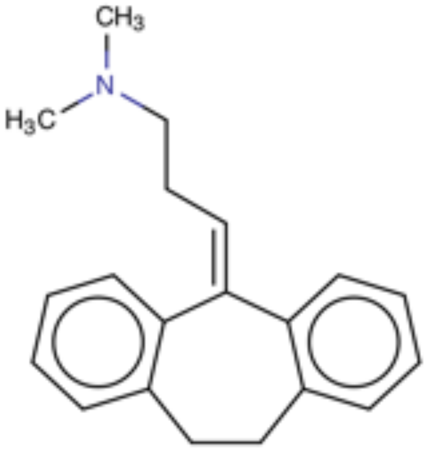
61



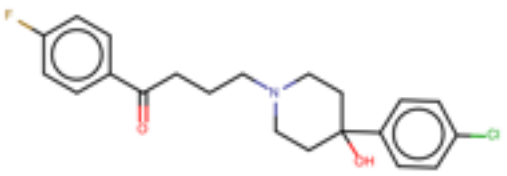
65



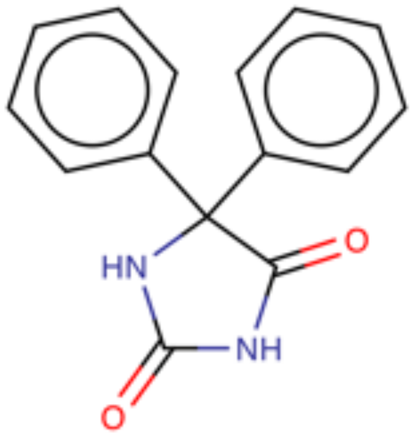
69



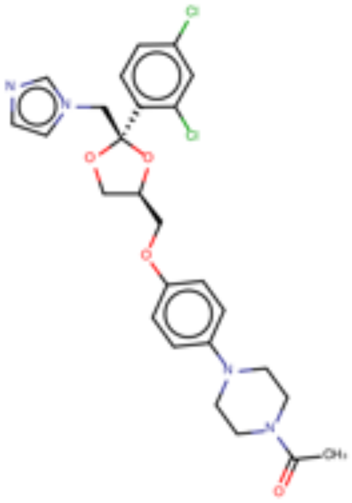
70



84

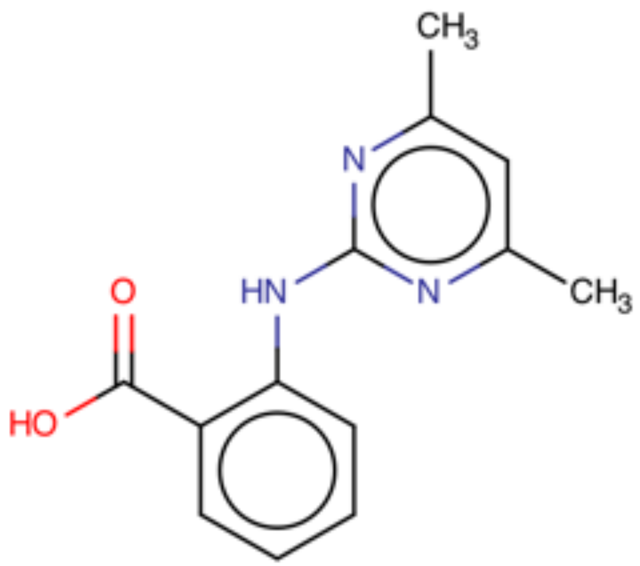


85

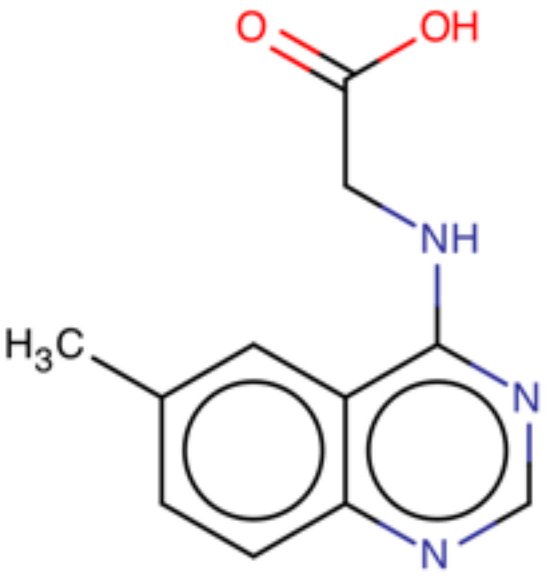


92

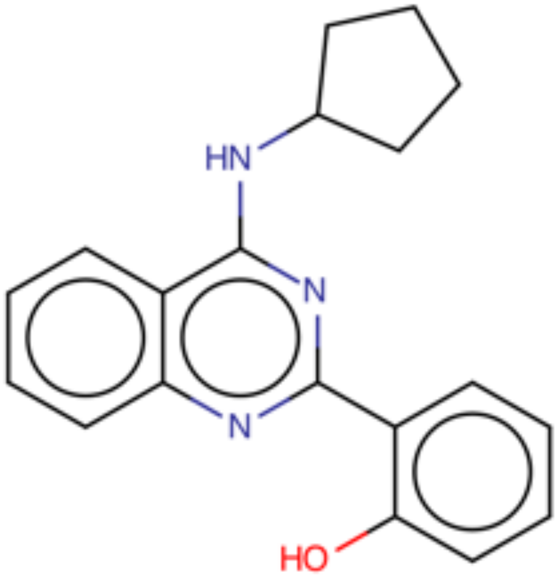
Test Set: p*K*$_a$ Baddies (less simple)



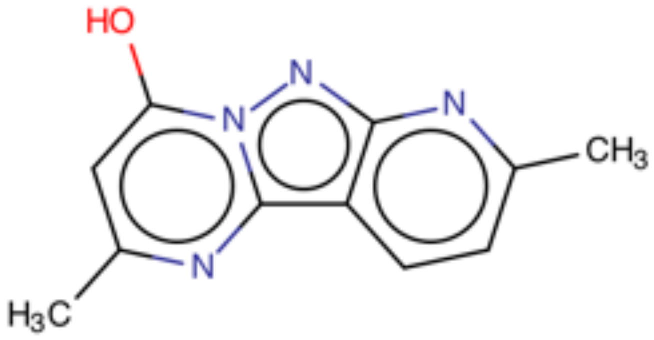
11



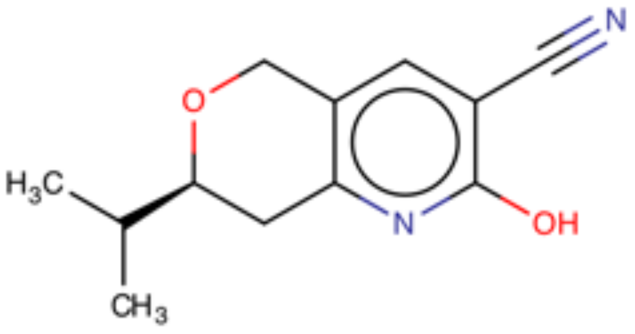
15



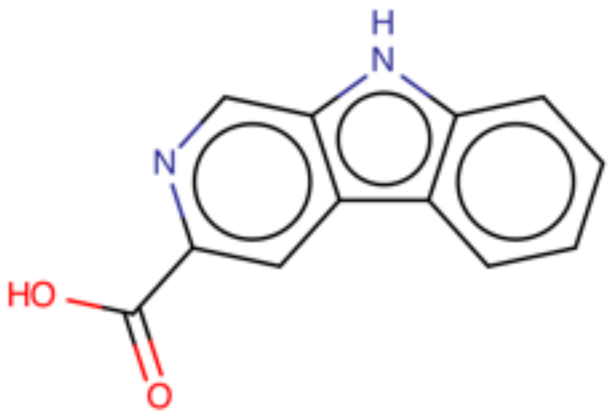
17



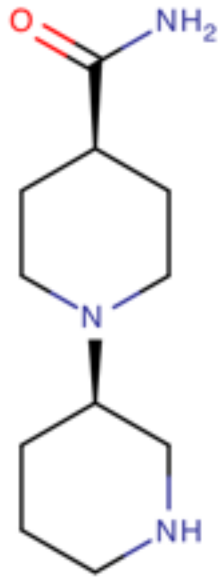
50



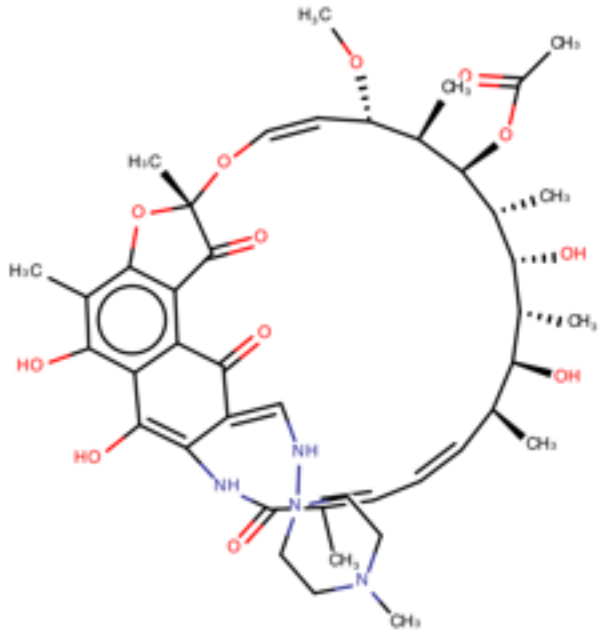
56



60

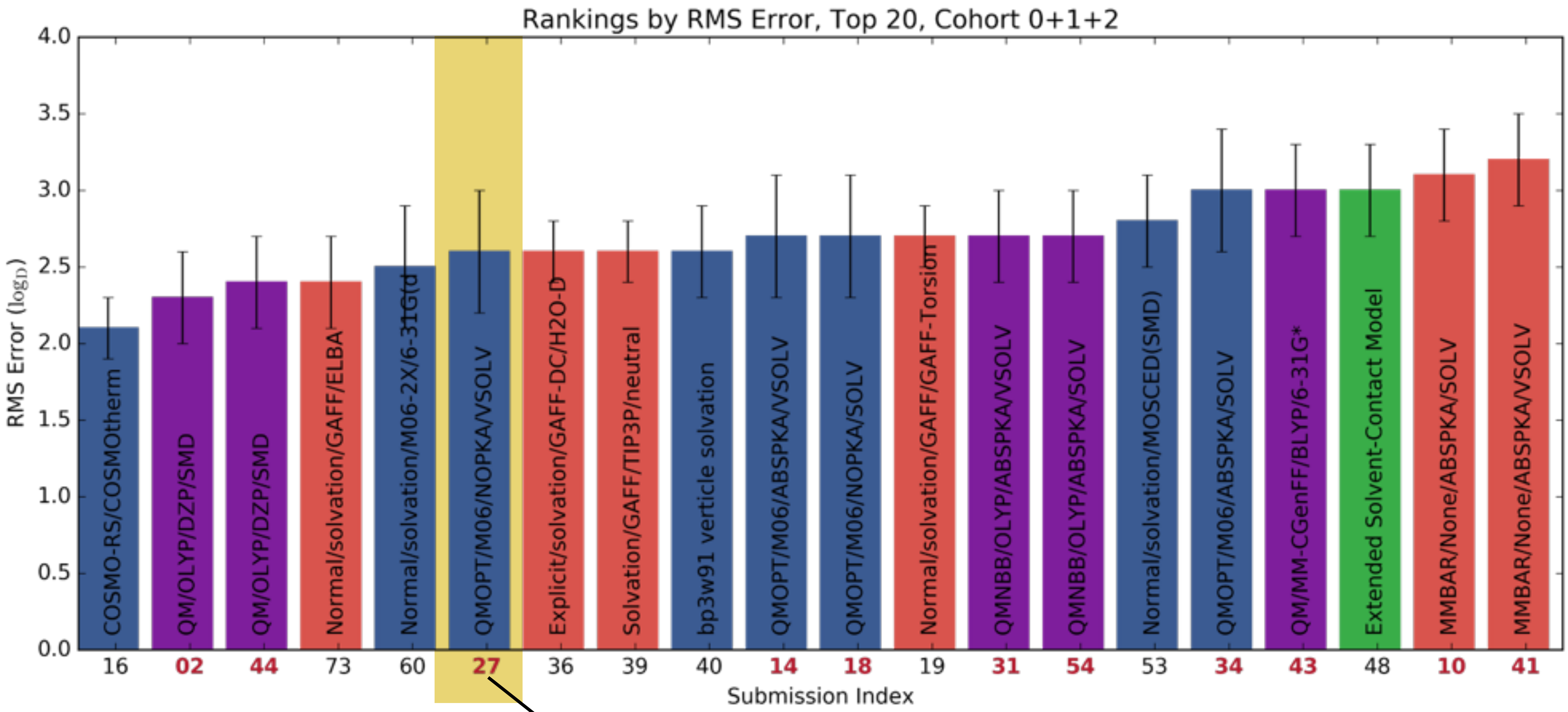


63



83

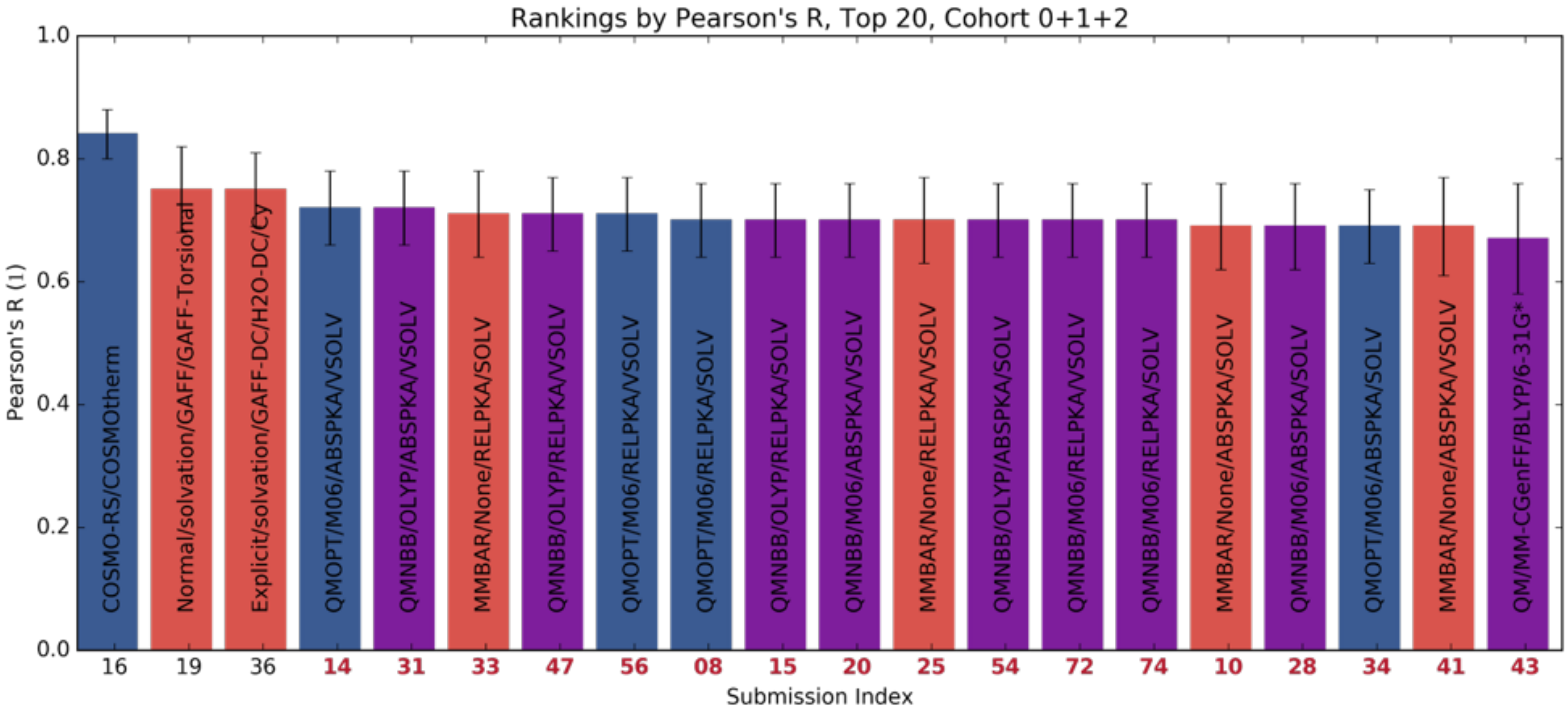
SAMPL5 Results



QM Control: M06-2X/6-31+G*/SMD

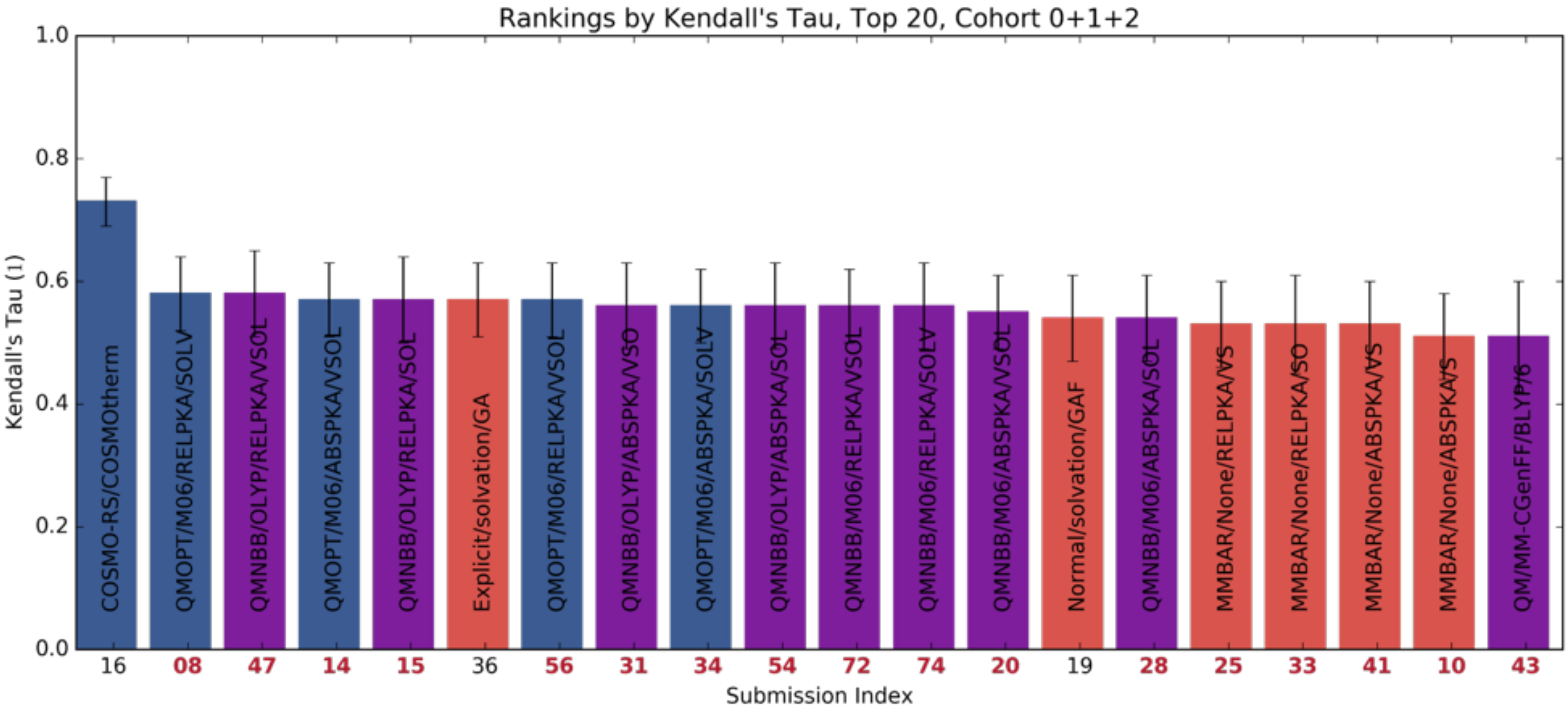
QM	Other
NBB	MM

SAMPL5 Results



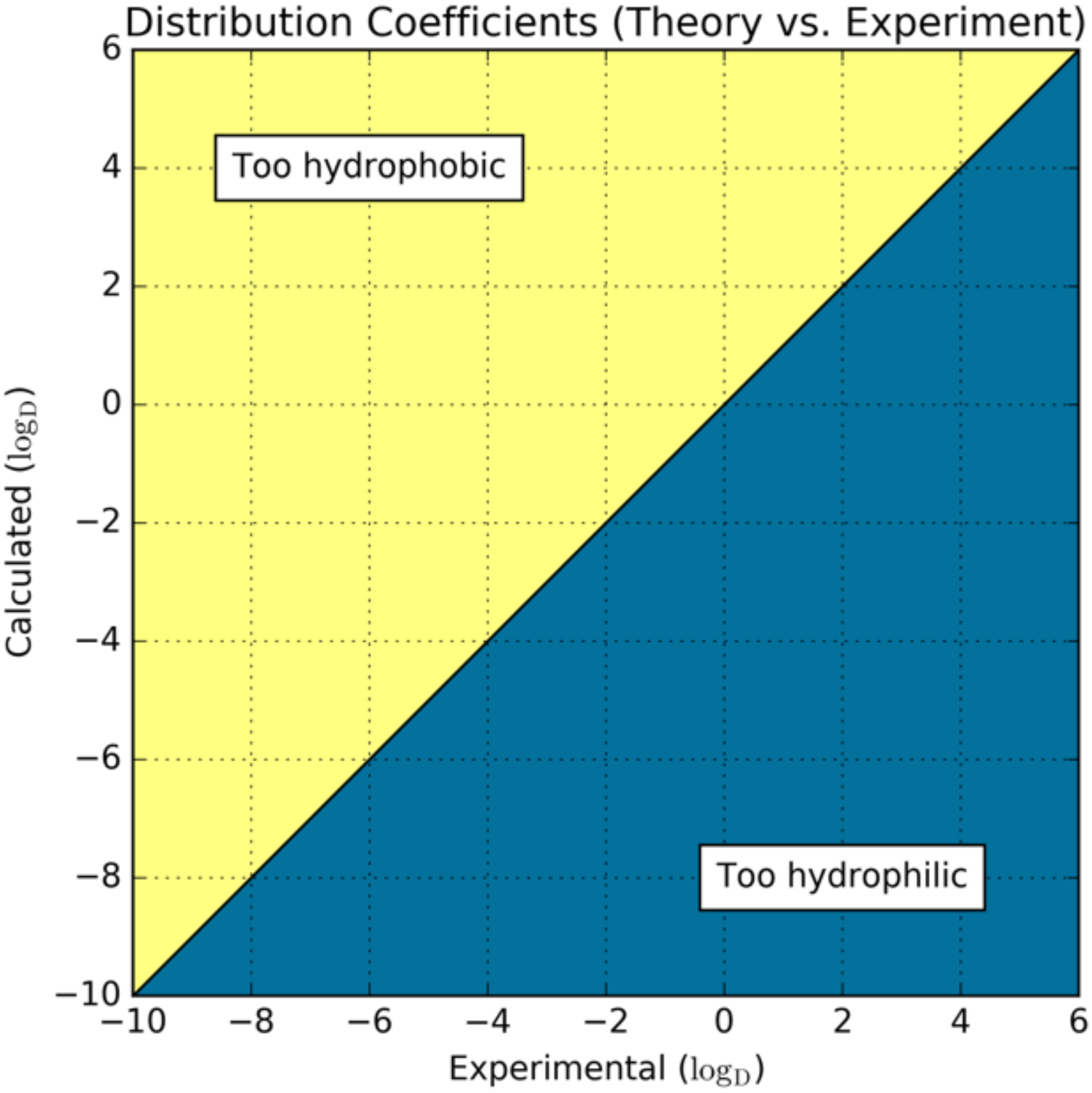
🟦 QM	🟩 Other
🟪 NBB	🟥 MM

SAMPL5 Results

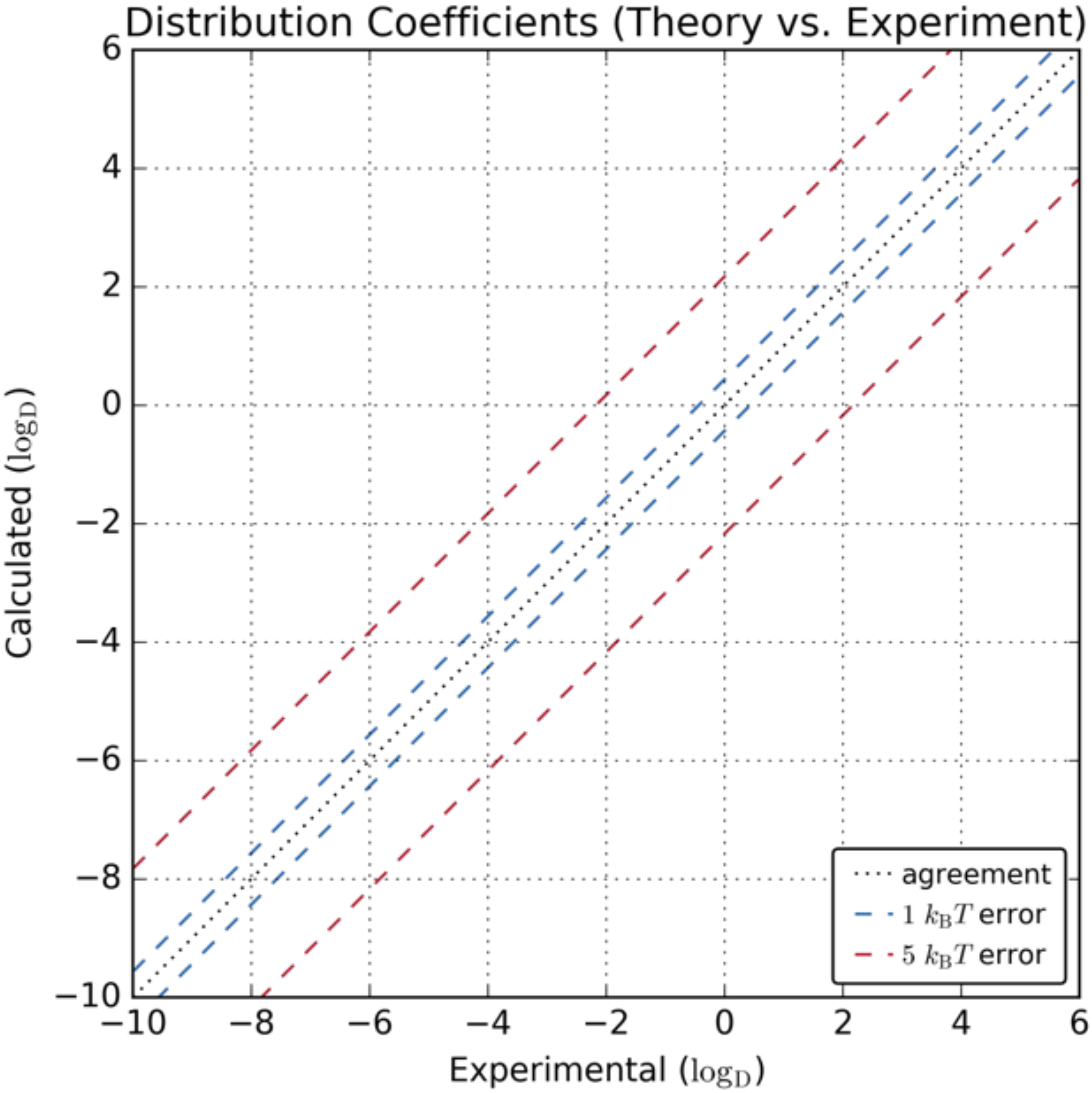


QM	Other
NBB	MM

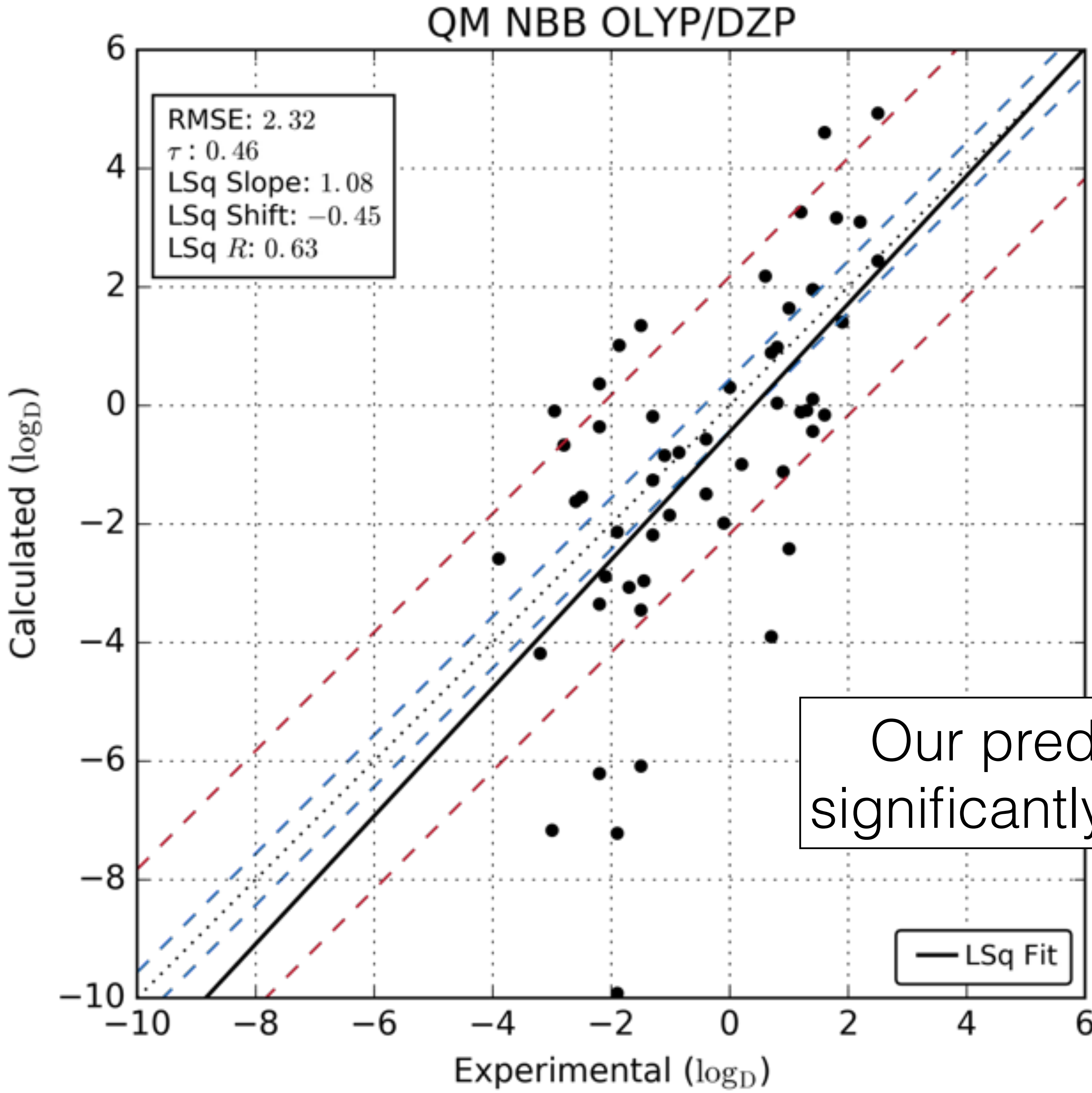
SAMPL5 Results



SAMPL5 Results

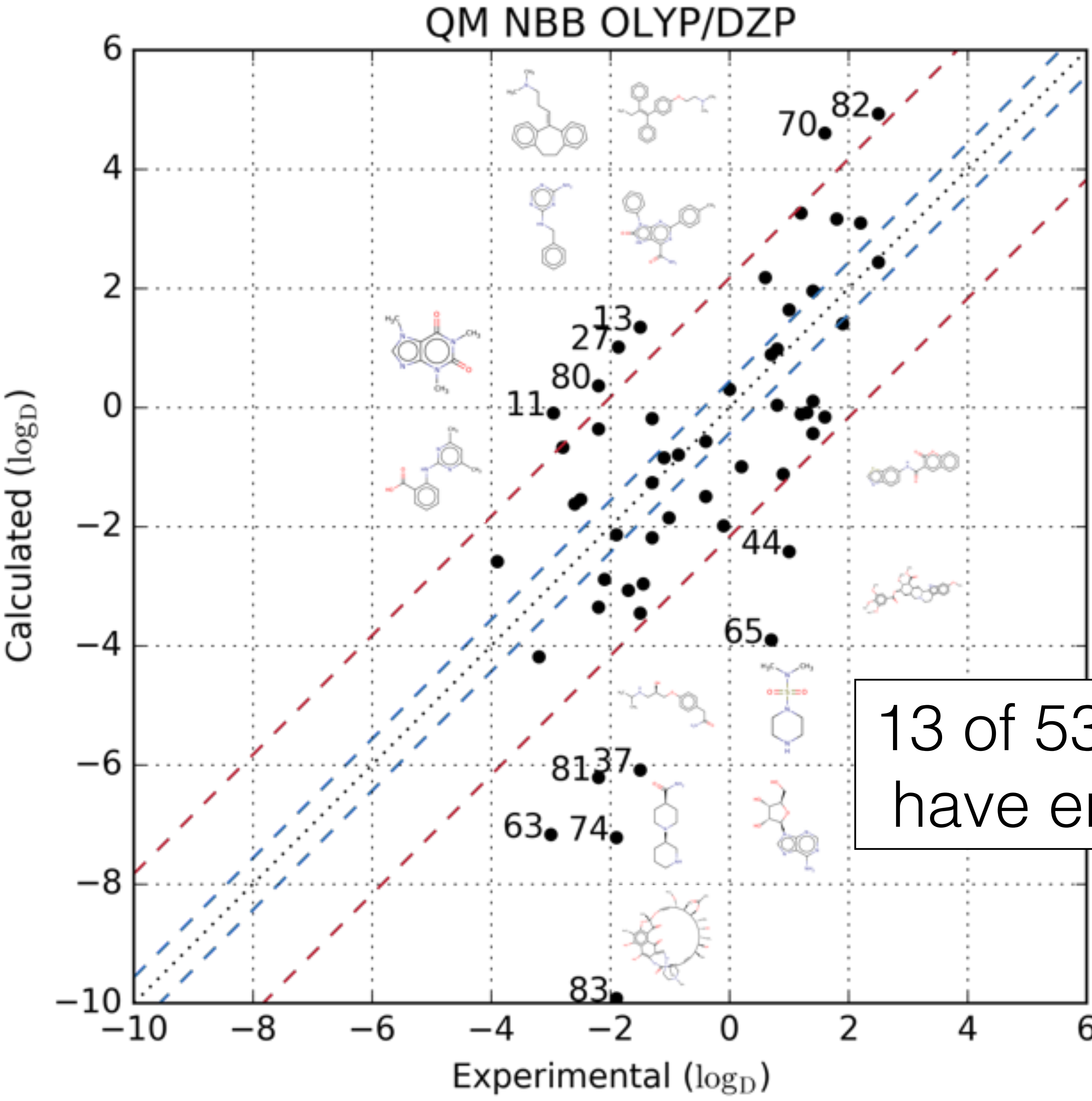


SAMPL5 Results



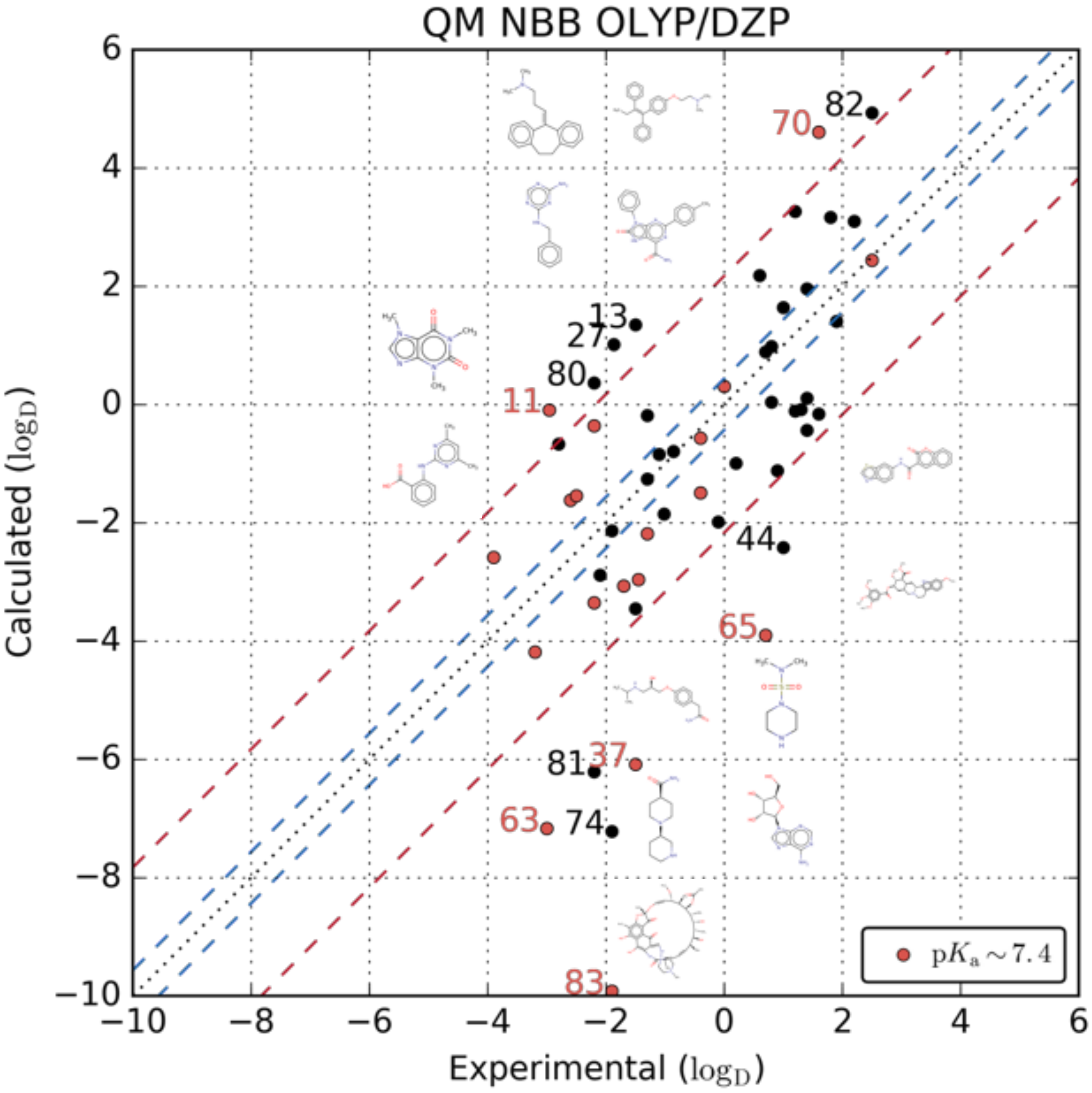
Our predictions are significantly hydrophilic.

SAMPL5 Results: The Baddies

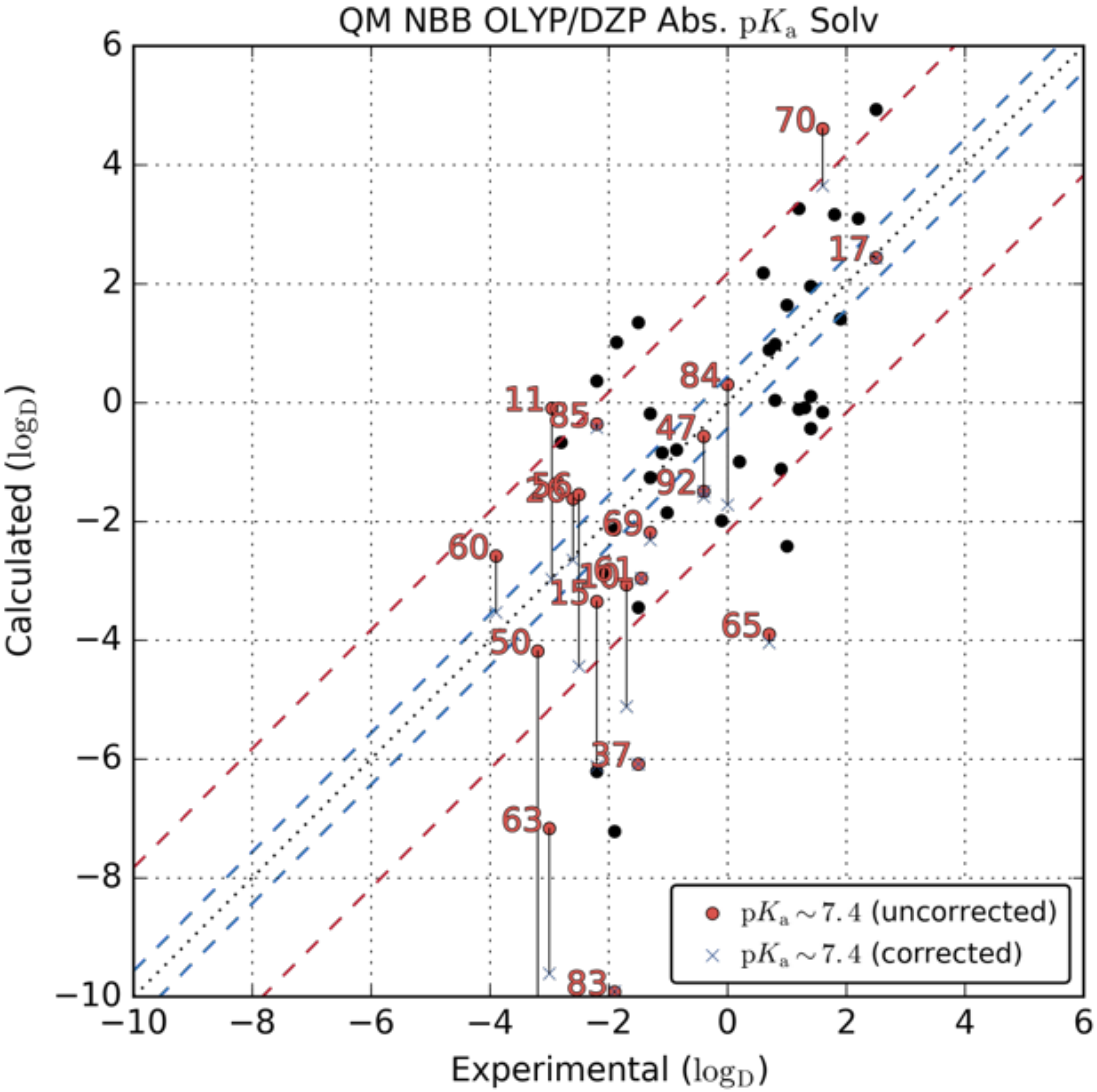


13 of 53 predictions have errors > 5 kT.

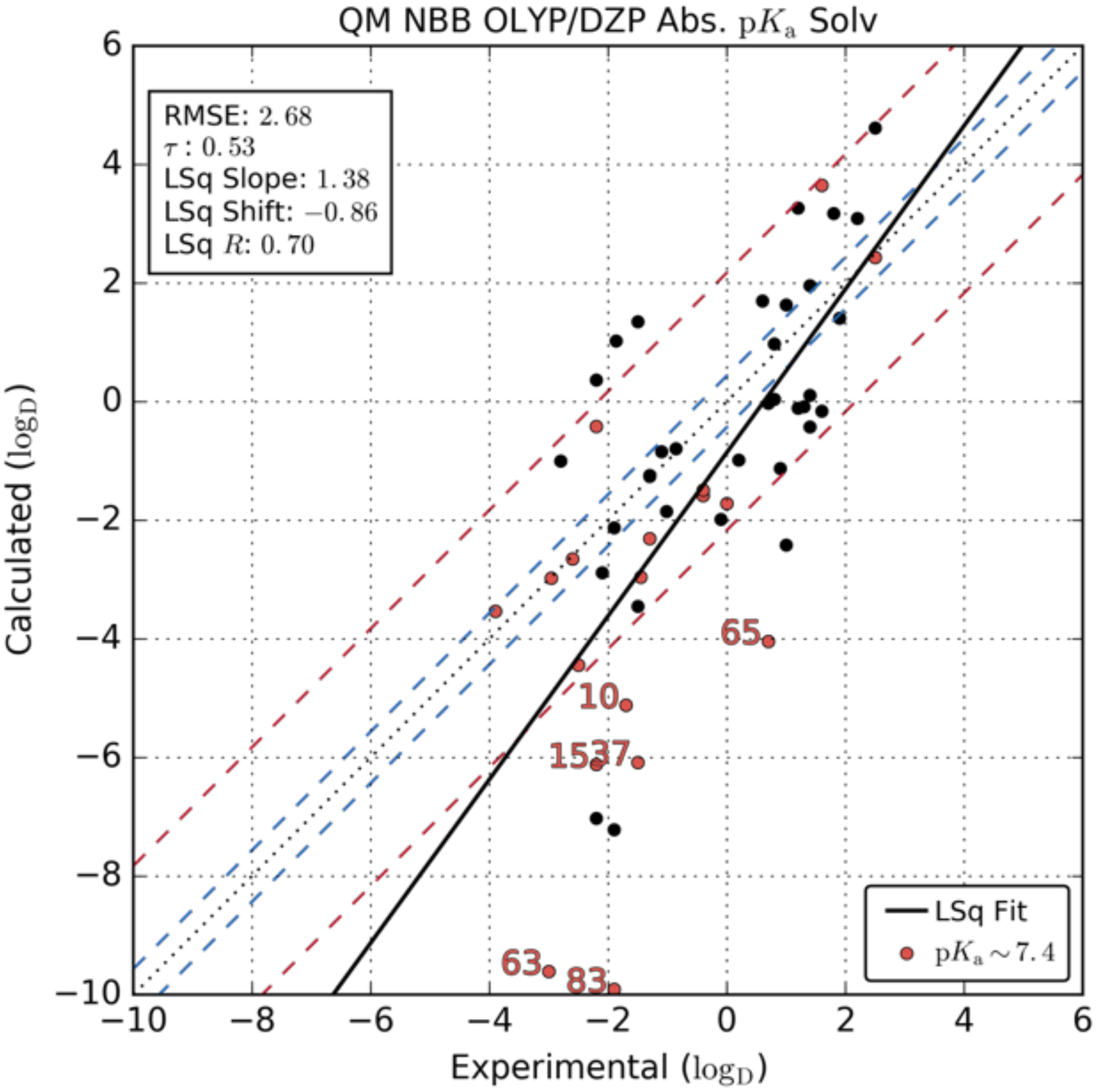
SAMPL5 Results: pK_a corrections



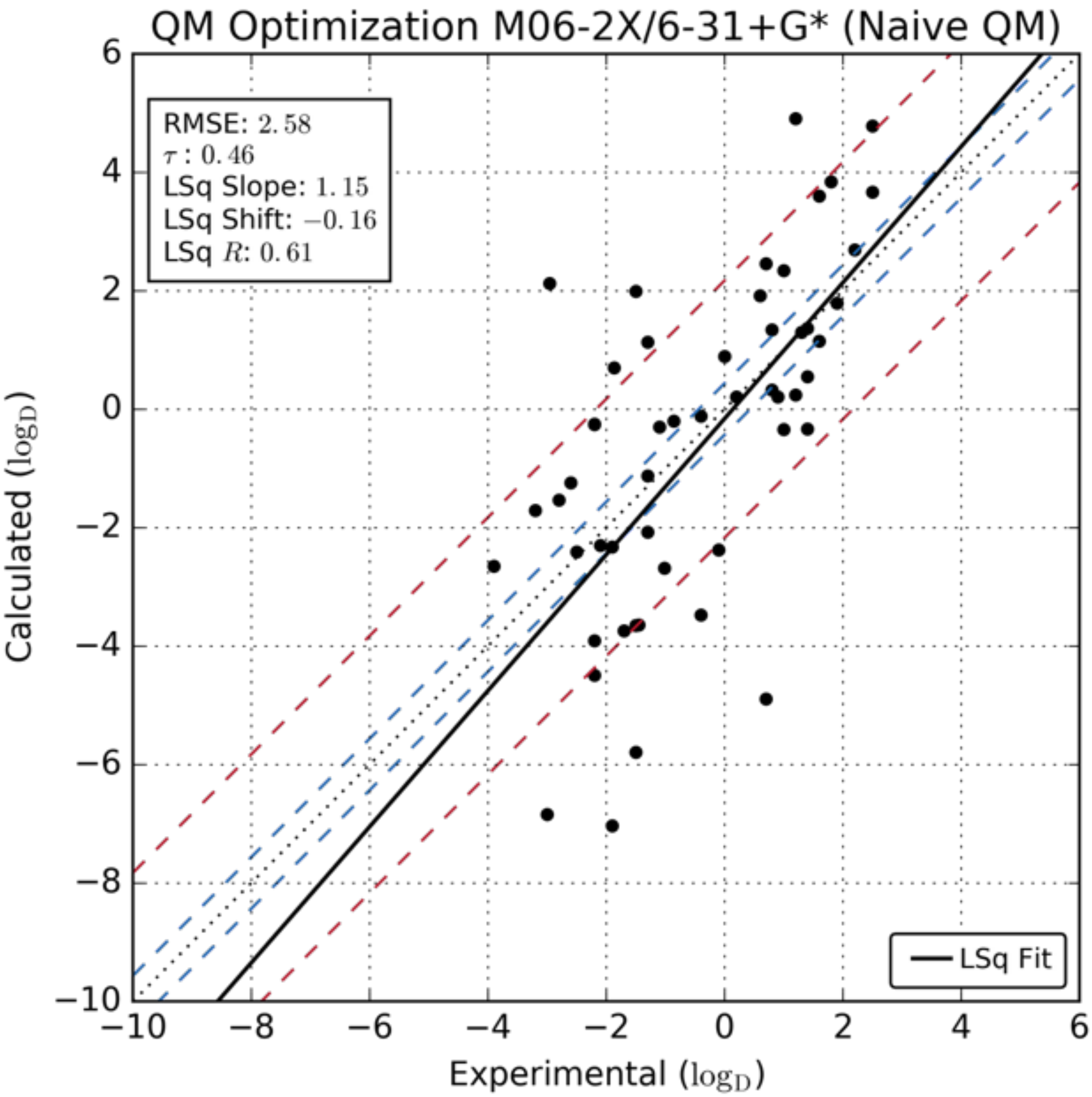
SAMPL5 Results: pK_a corrections



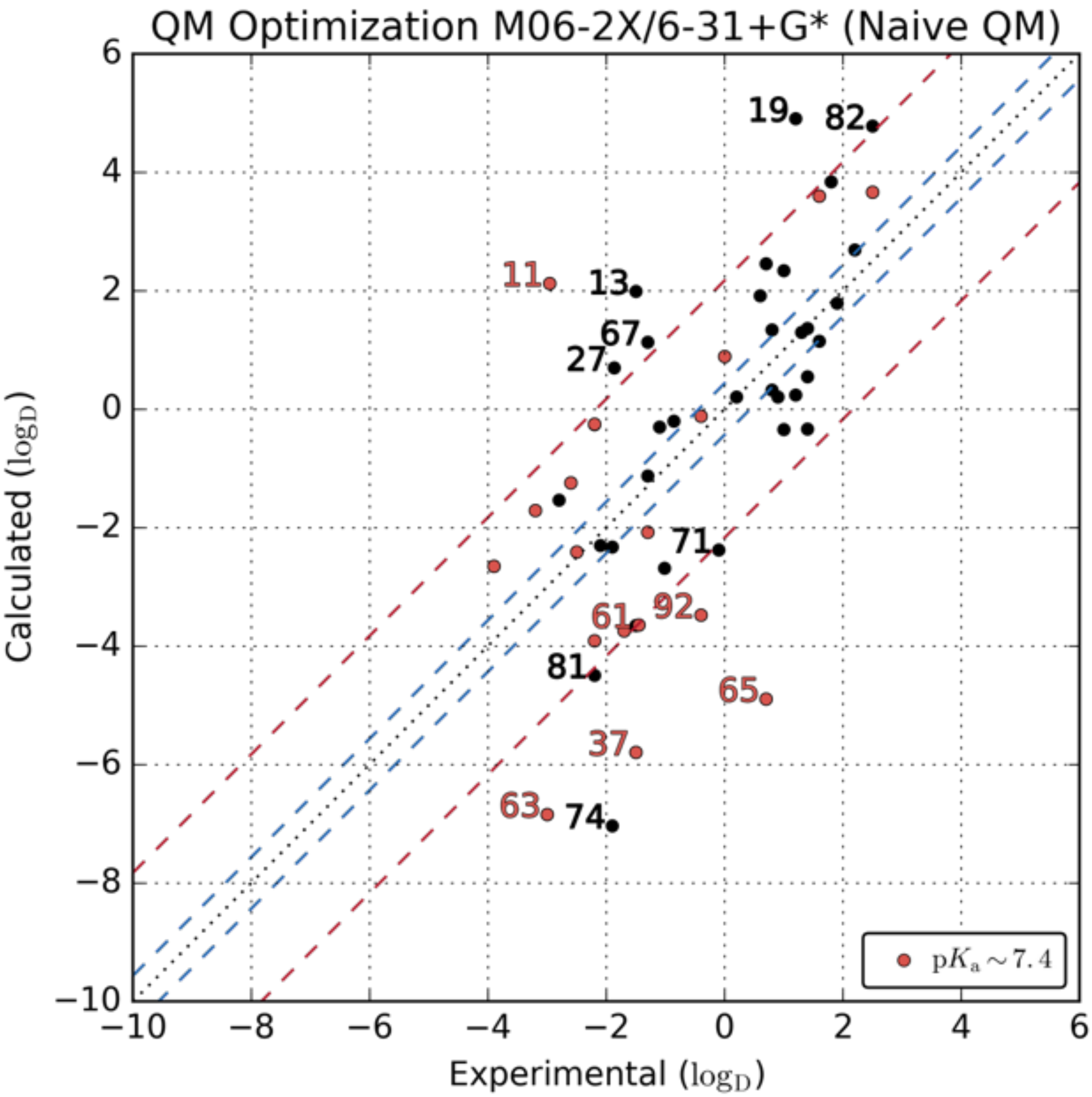
SAMPL5 Results: pK_a corrections



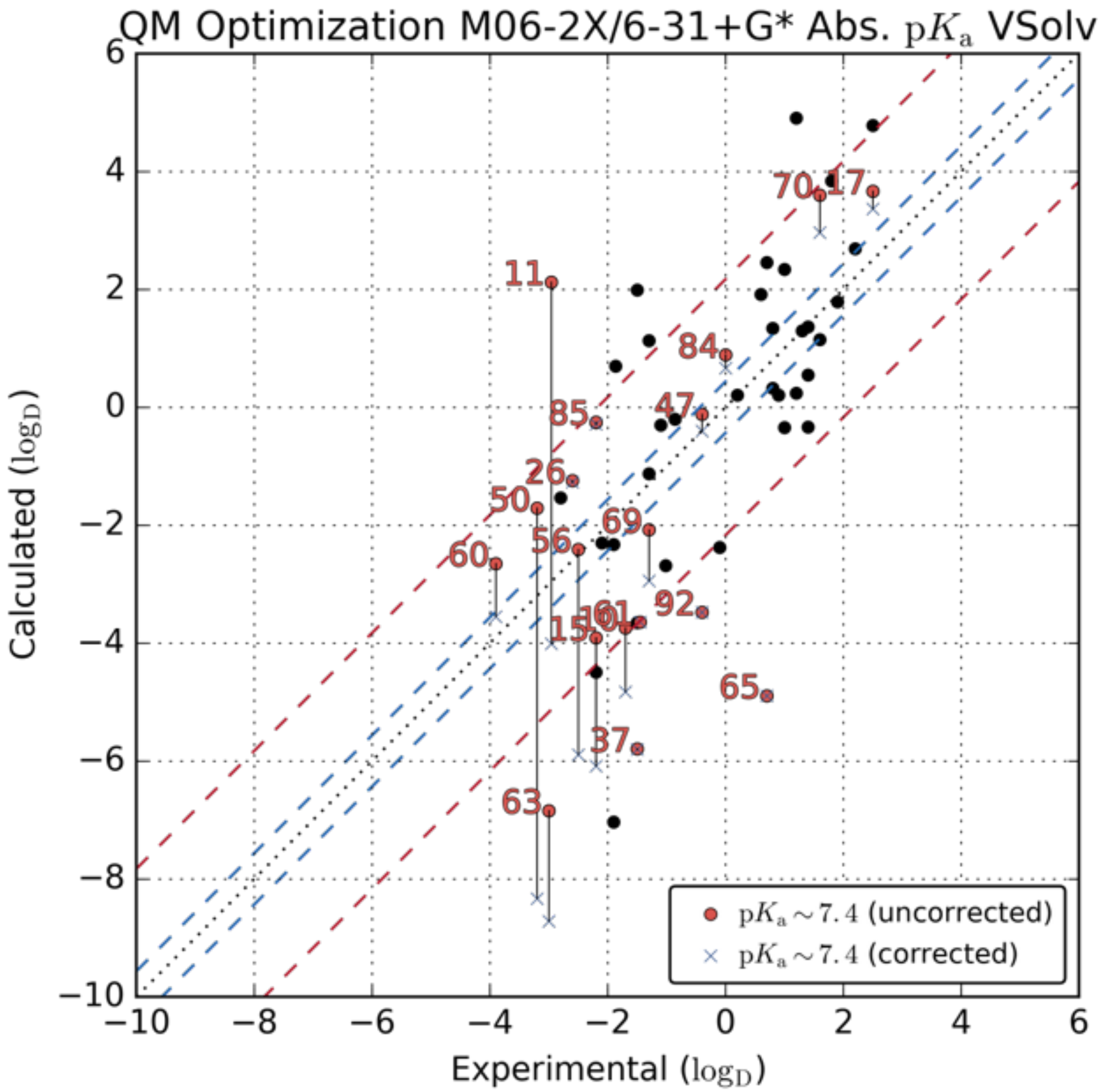
SAMPL5 Results



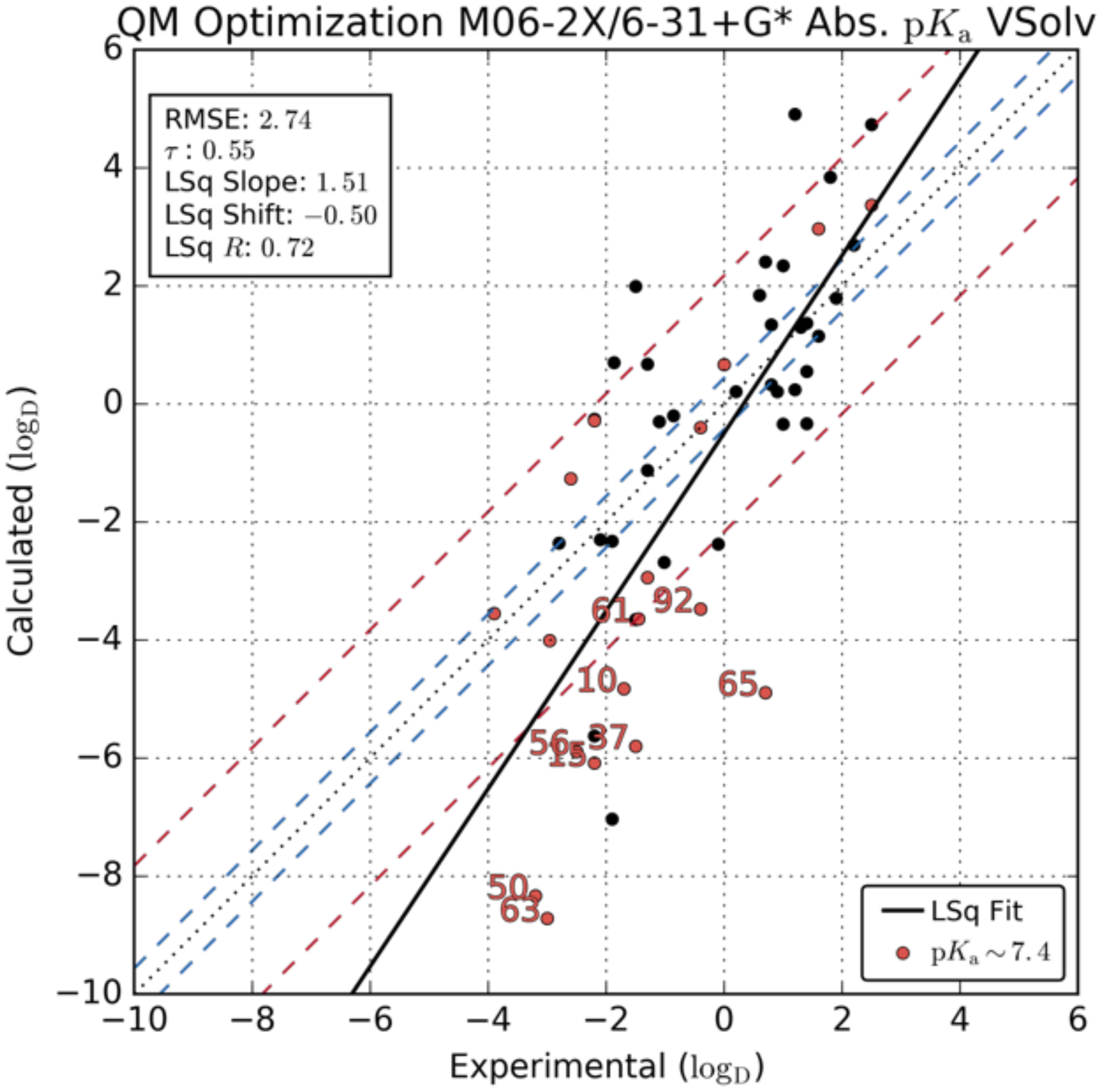
SAMPL5 Results



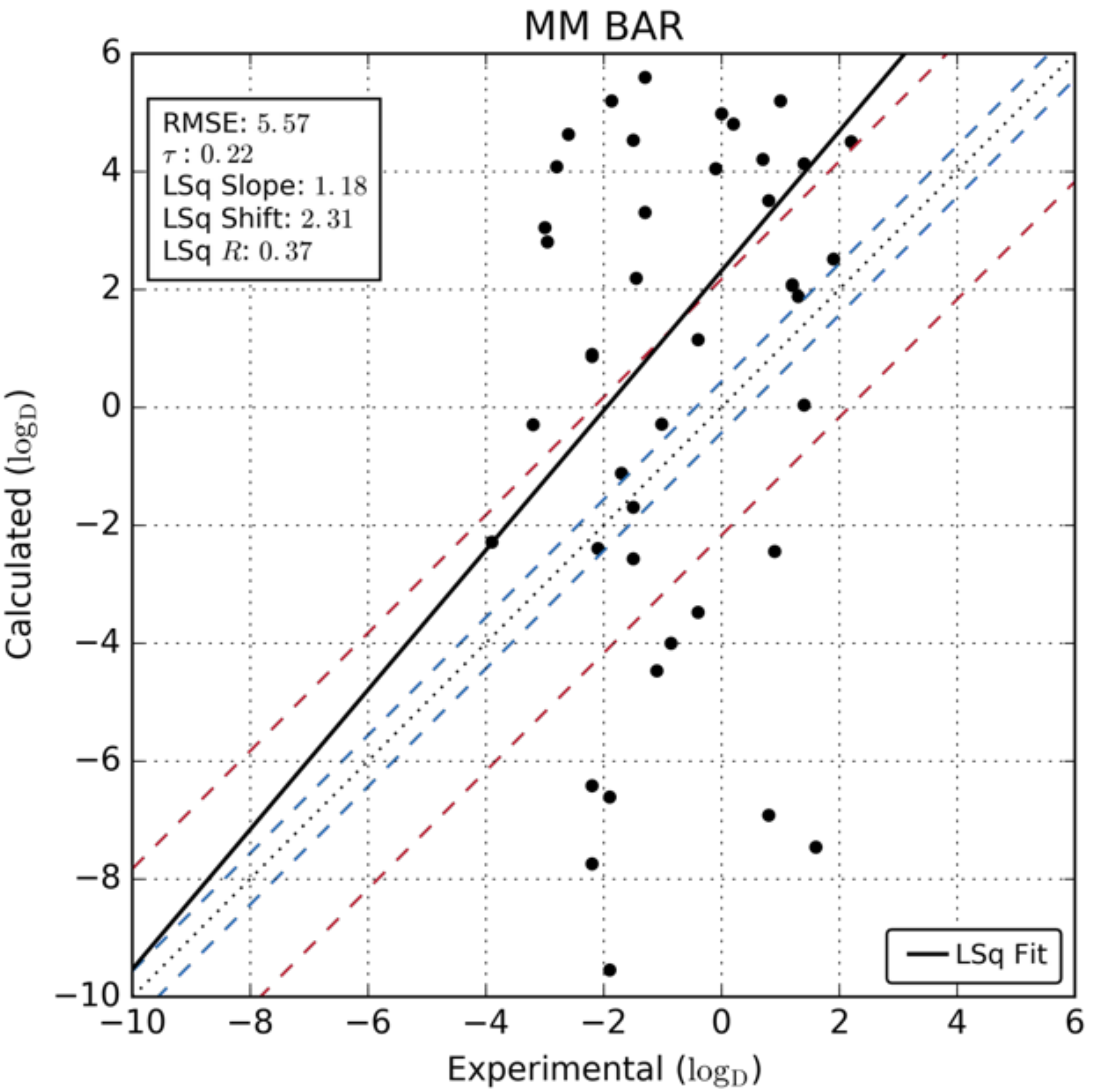
SAMPL5 Results



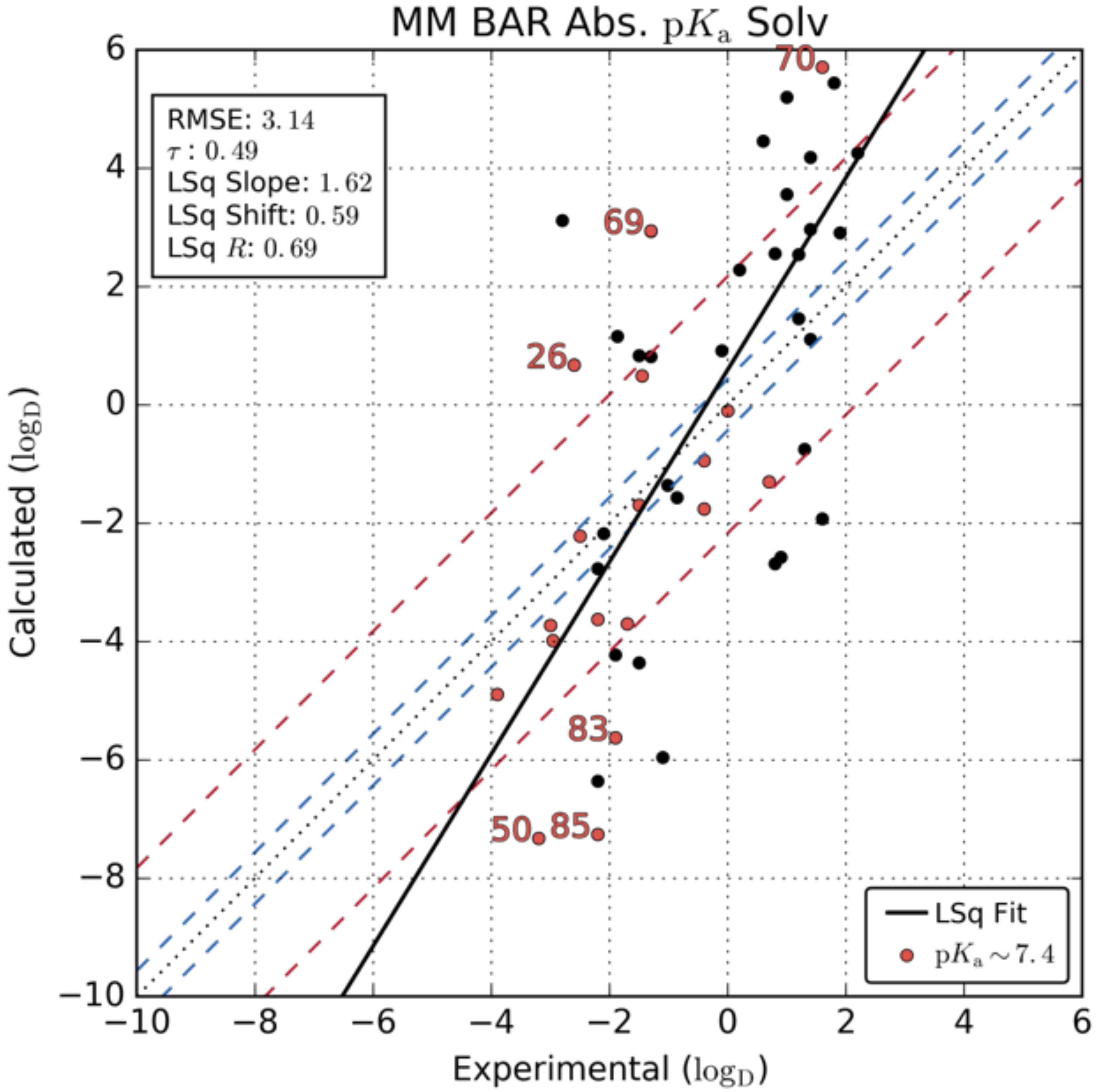
SAMPL5 Results



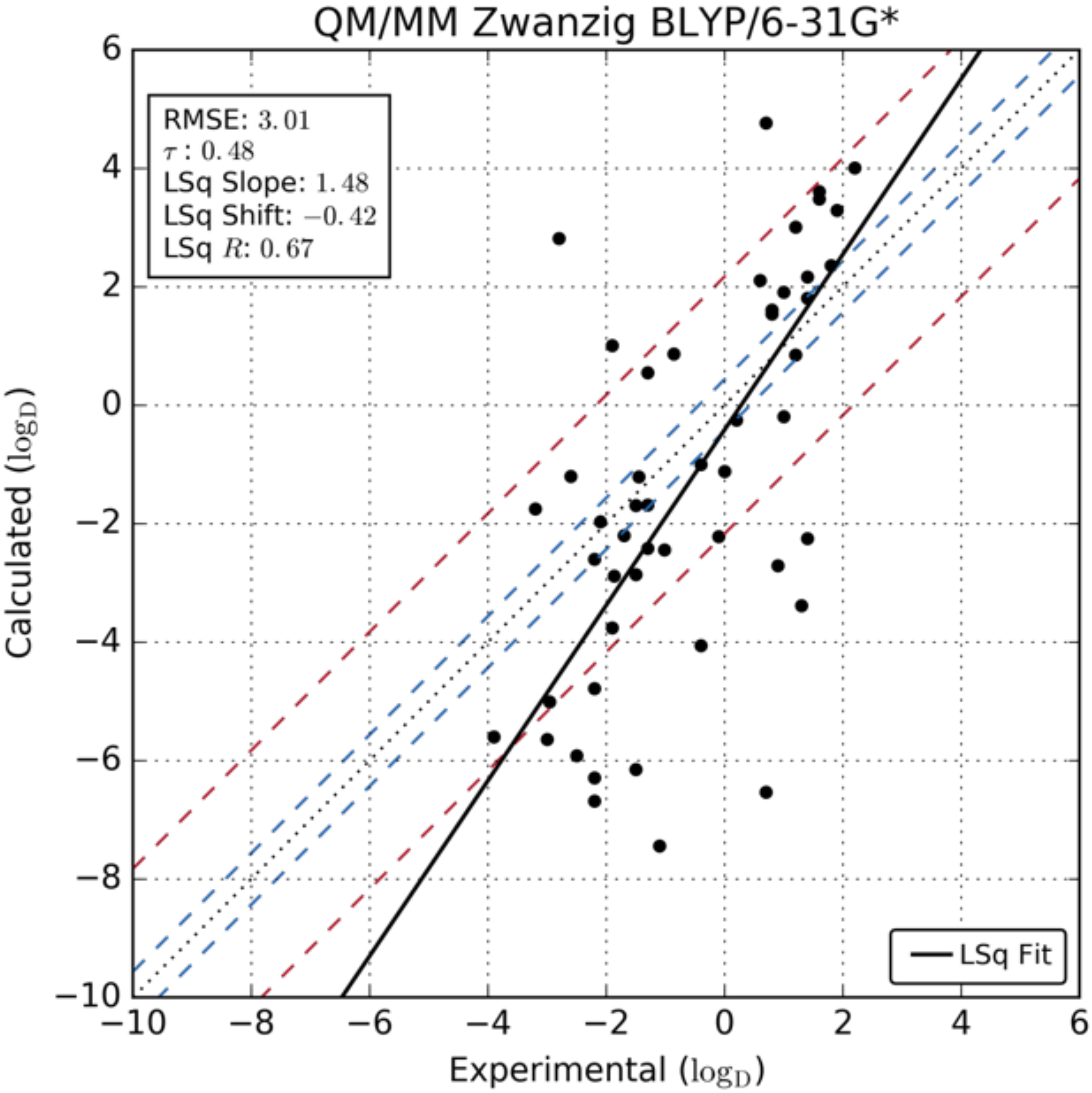
SAMPL5 Results



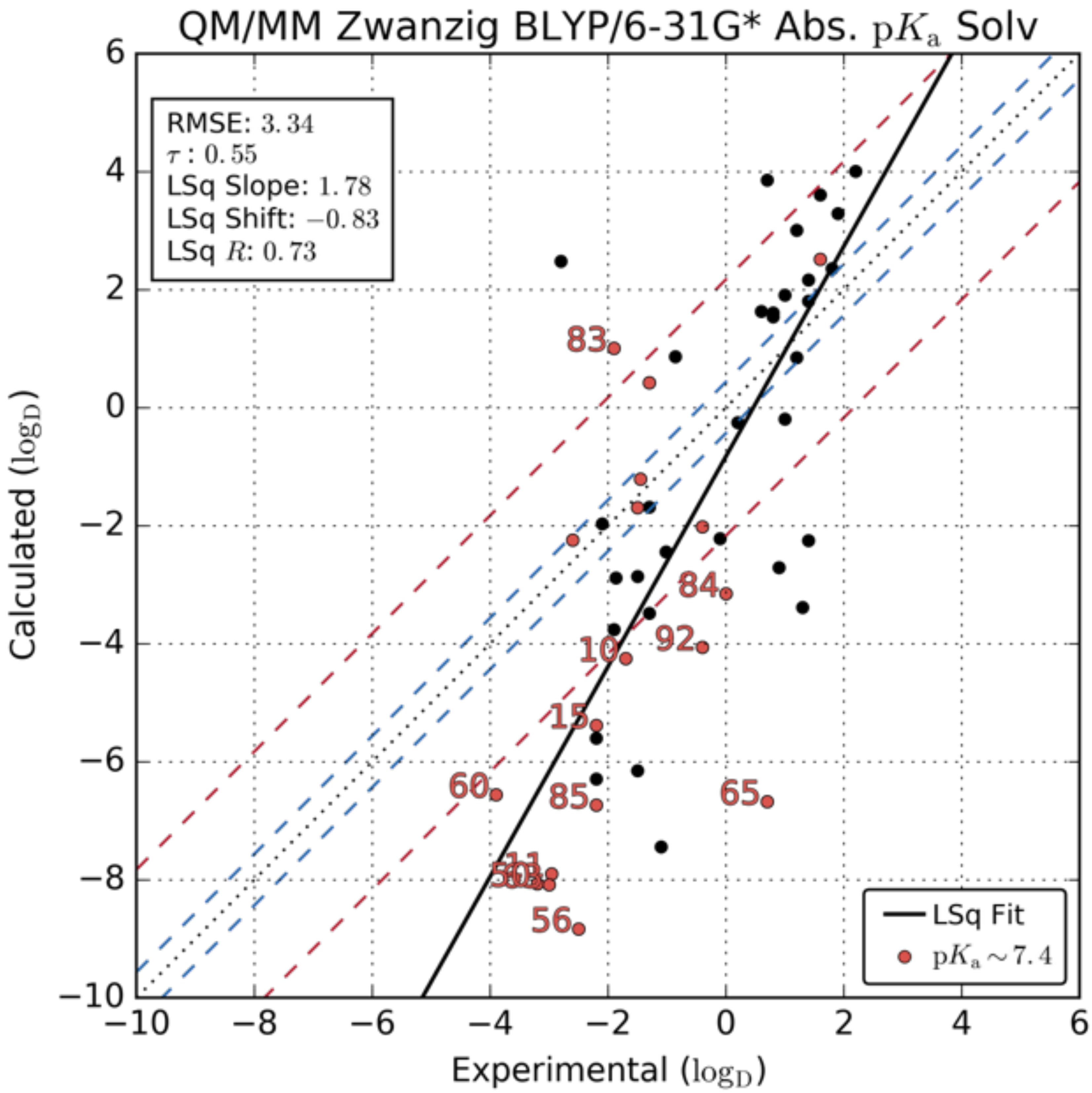
SAMPL5 Results



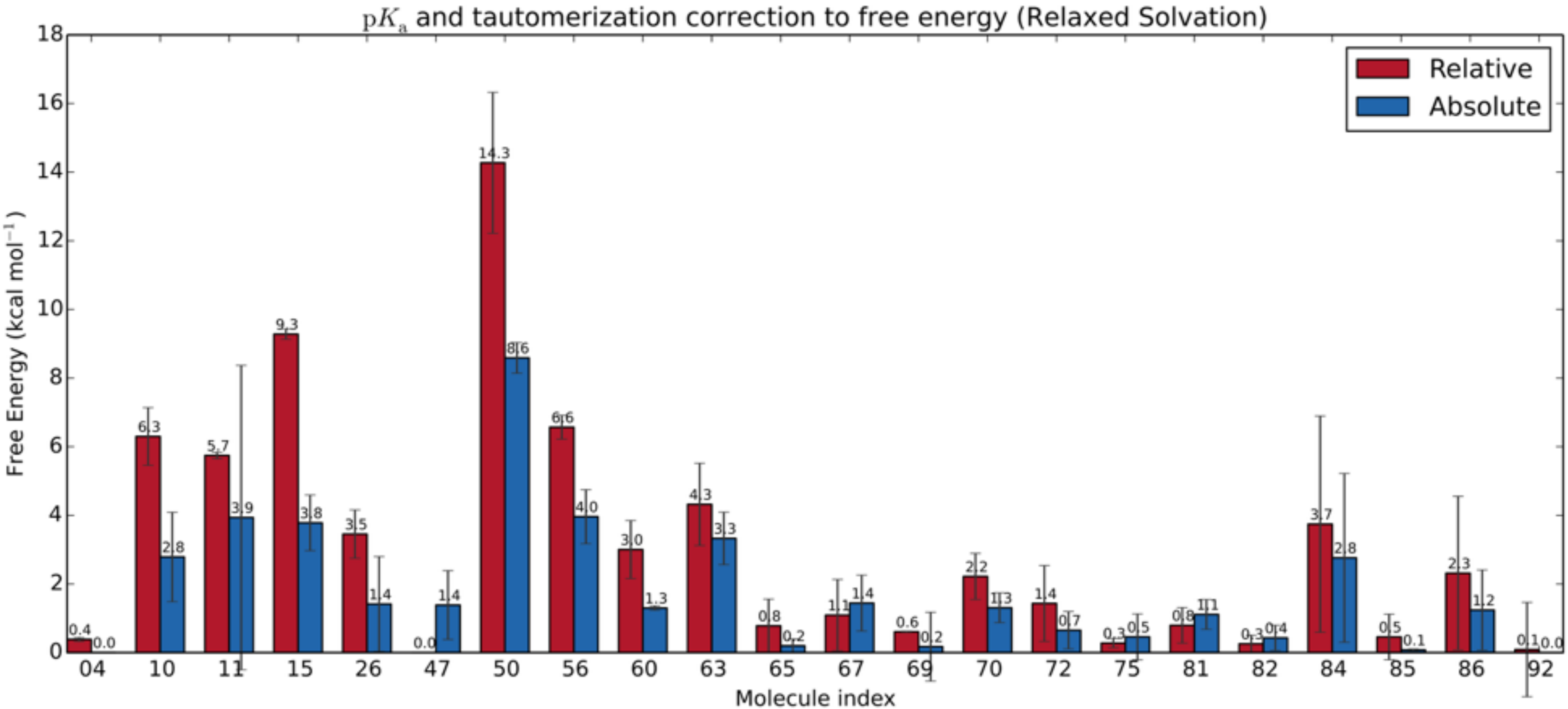
SAMPL5 Results



SAMPL5 Results



SAMPL5 Results: pK_a corrections



Large differences between pK_a methods might indicate tautomerization issues.

SAMPL5 Results: Conclusions

- Predicting log D values is *difficult*.
- Lessons learned from SAMPL4 have carried over (choice of density functional and basis set).
- NBB QM calculations with implicit solvent are among the best options (RMSE rank 2nd) but have poor correlation.
- Our predictions are too hydrophilic, we ignored the wetness in cyclohexane
- Accounting for tautomers is very important, and universally improves our correlation (but reduces RMSE).