MIAMI UNIVERSITY
OXFORD, OH • EST. 1809

log P Predictions Using SMx or LSER With Training Data from DrugBank.ca

Jonathan A. Ouimet, Rachel C. Ollier, and Andrew S. Paluch

Department of Chemical, Paper, and Biomedical Engineering
Miami University

PaluchAS@MiamiOH.edu

SAMPL6 Virtual Workshop

May 16, 2019

Our Entries

- SM12-Solvation-Trained (7/91)
- SM12-Solvation
- SM8-Solvation-Trained
- SM8-Solvation
- SMD-Solvation-Trained
- SMD-Solvation

- GC-LSER
- ISIDA-LSER
- UFZ-LSER

The Idea

The standard approach:

$$\log P = -\frac{1}{\ln(10)}\left(\frac{\Delta G^{\mathrm{solv}}_{\mathrm{oct}}}{RT} - \frac{\Delta G^{\mathrm{solv}}_{\mathrm{water}}}{RT}\right)$$

Which assumes:

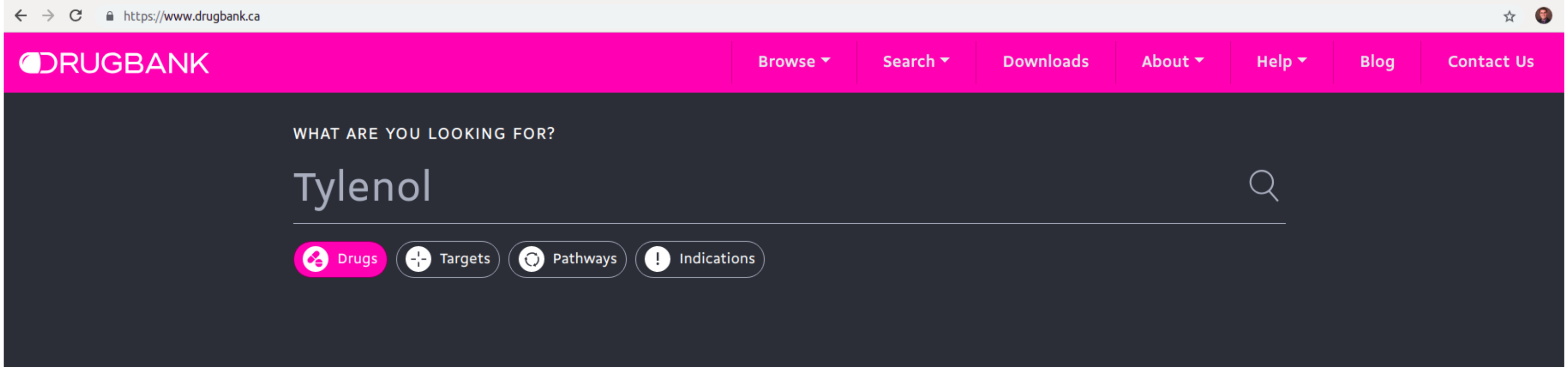
1) Solute is infinitely dilute (i.e., no solute-solute interactions)
2) Water and octanol are immiscible

What if instead we used:

$$\log P = -a\frac{\Delta G^{\mathrm{solv}}_{\mathrm{oct}}}{RT} + b\frac{\Delta G^{\mathrm{solv}}_{\mathrm{water}}}{RT} + c$$

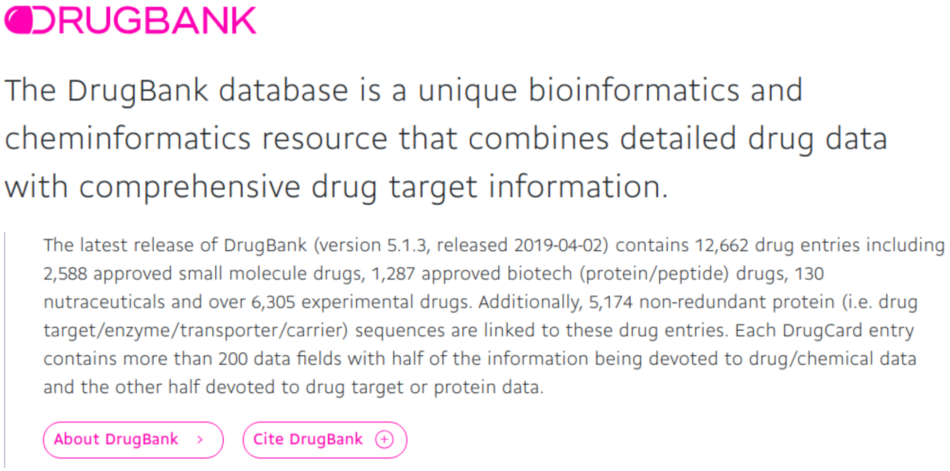
Training Set

- We sought a training set representative of the SAMPL6 molecules for which predictions were to be made.



WHAT ARE YOU LOOKING FOR?

Tylenol



The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.

The latest release of DrugBank (version 5.1.3, released 2019-04-02) contains 12,662 drug entries including 2,588 approved small molecule drugs, 1,287 approved biotech (protein/peptide) drugs, 130 nutraceuticals and over 6,305 experimental drugs. Additionally, 5,174 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

Training Set

- We sought a training set representative of the SAMPL6 molecules for which predictions were to be made.
  - 6 of the molecules had the 4-amino quinzaoline scaffold
  - all the molecules had multiple aromatic rings within their structure
- Selected molecules that
  - Had the 4-amino quinzaoline scaffold
  - Experimental log *P* values were available
  - size and molecular weight were considered to ensure that the selected molecules were representative of the SAMPL6 molecules
  - Include halogens and electronegative elements
- Final training set of 100 molecules from DrugBank.ca

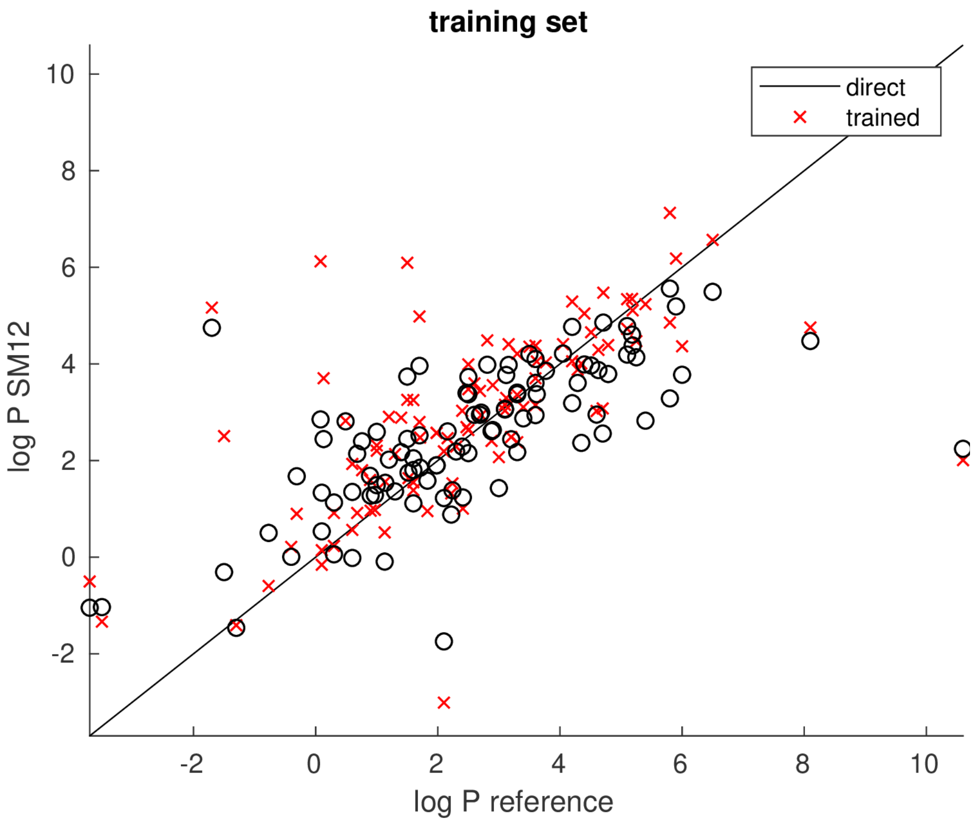
Test/Validation Set

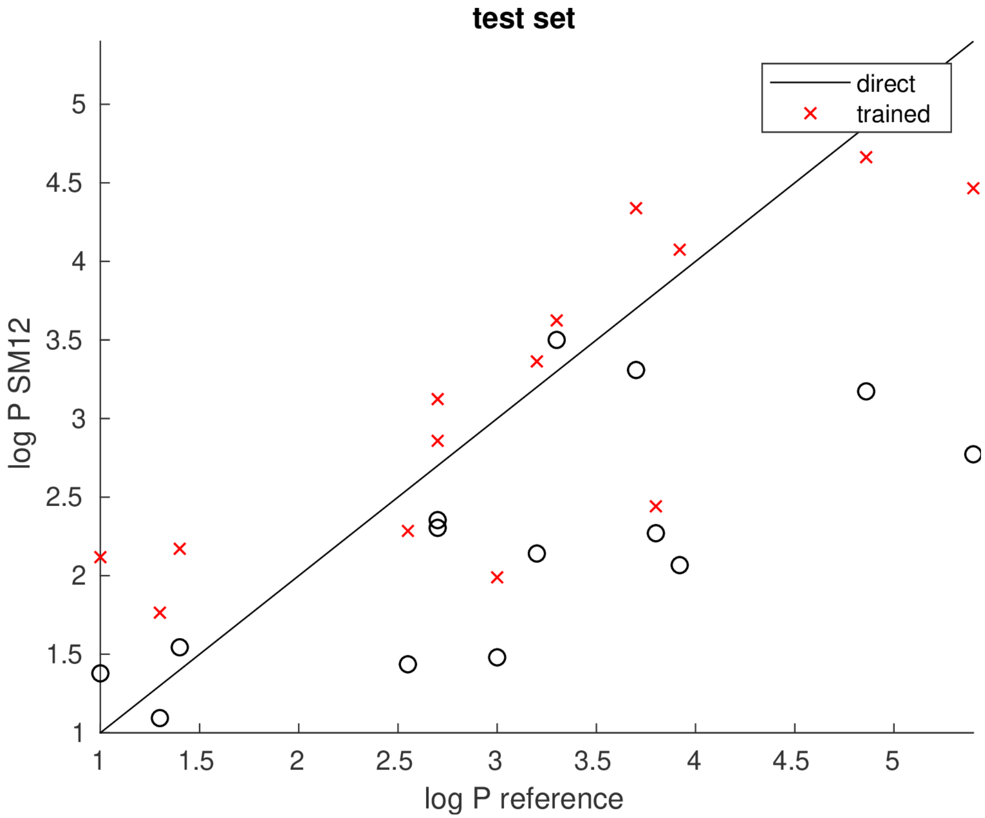
- Performed a similarity in the DrugBank.ca database
  - Similarity score greater than 0.5 with the SAMPL6 structures
  - Also included “approved” drugs with know log *P* values
- Test set of 14 molecules
- Used to estimate model accuracy

Calculations

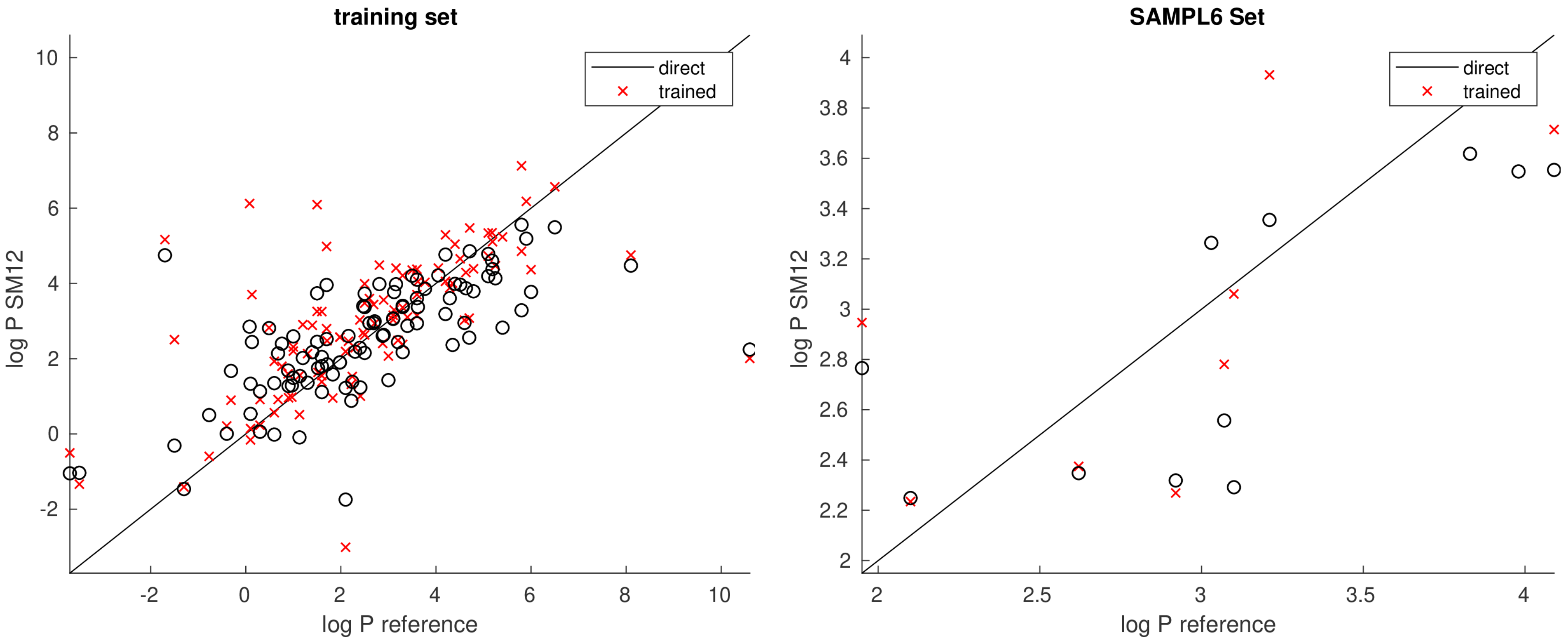
- Start with SMILES from DrugBank.ca and SAMPL6
- Open Babel using GAFF
  - Generate 3-D structures and then performed systematic conformation search to find lowest energy conformer
- QChem 5.1.2
  - Geometry optimization at the M06-2X/cc-pVDZ level of theory/basis set
  - Single point energy energy calculations in the SM12, SM8, and SMD continuum solvent models for 1-octanol and water at the M06-2X/6-31G(d) level of theory/basis set

Top Performing SM12



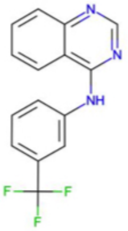


Top Performing SM12

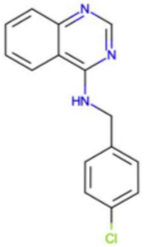


Possible sources of error within our control

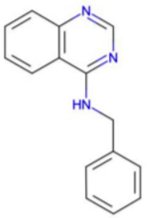
- Conformation
- Theory/basis set
- More descriptors and/or machine learning in place of multi-linear regression
- How to treat HCl
  - Performing calculations without HCl for publication



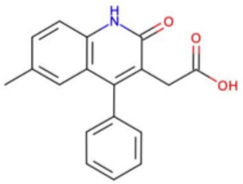
SM02



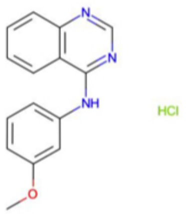
SM04



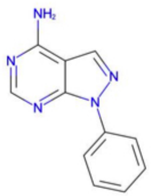
SM07



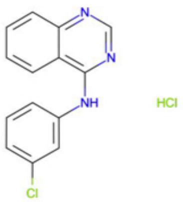
SM08



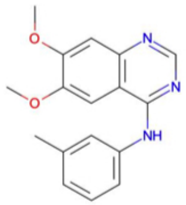
SM09



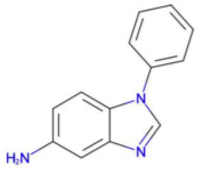
SM11



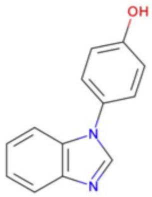
SM12



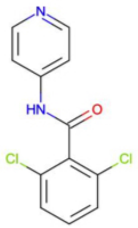
SM13



SM14



SM15



SM16

Some Lessons Learned

- Computed solvation free energies are sensitive to choice of SM12, SM8, or SMD.
  - The results are expected to be sensitive to theory/basis set too.
  - And geometry.
- Including “training” data did effect the results. But it is unclear how significant this is.
  - Could also investigate choice of training data further.
  - DrugBank.ca was chosen due to its size and it is freely available.
- HCl

Acknowledgements

- Ohio Supercomputer Center

PaluchAS@MiamiOH.edu