ESTIMATE WATER SOLUBILITY

Christos Kannas

University of Cyprus

Department of Computer Science

Outline

- Introduction
- Related Work
  - ESOL
- RDKit based Implementation
- Results
  - Correlation Table & Chart
- Conclusion

University of Cyprus
Department of Computer Science

RDKit
Open-Source Cheminformatics and Machine Learning

SEVENTH FRAMEWORK PROGRAMME

LiSIs
Life Sciences Informatics System

Introduction

- Need to estimate the solubility of molecules in:
  - DMSO (CS(=O)C), and
  - Water.

- Predictive Models for DMSO and Water Solubility.

Related Work

Related Work

- J. S. Delaney, “ESOL: Estimating Aqueous Solubility Directly from Molecular Structure,” *Journal of Chemical Information and Modeling*, vol. 44, no. 3, pp. 1000–1005, May 2004.

University of Cyprus
Department of Computer Science

Related Work: ESOL

- ESOL – Estimated SOLubility
- Linear Regression Model
- 8 Molecular Properties (Initially)

- Preeminent Method: General Solubility Equation (GSE), logP and melting point (T_m)

ESOL: Molecular Properties (Initial) 1/3

- **clogP** – Daylight CLOGP v4.72
- **MolWeight**
- **RotBonds** – Rotatable Bonds, Daylight SMARTS structures define rotatable bonds

ESOL: Molecular Properties (Initial) 2/3

- **Aromatic Proportion (AromProp)** – The proportion of heavy atoms in the molecule that are in an aromatic ring. Daylight SMARTS ([a]) aromatic atoms.

- **Non-Carbon Proportion** – The proportion of heavy atoms in a molecule that are not carbon. Daylight SMARTS ([!#6])

ESOL: Molecular Properties (Initial) 3/3

- **H-bond Donors**
- **H-bond Acceptors**
- **Polar Surface Area** – Peter Ertl's Polar Surface Area

ESOL: Methodology

- Multiple Linear Regression
- Significance of each parameter based in terms of its absolute t-statistic.

ESOL: Train Dataset

- Training Set: 2874 molecules
  - Small – Low MolWeight organic compounds
  - Medium – Pesticide products, MolWeight 200-300
  - Large – Sygenta compounds, MolWeight 300-400

ESOL: Results

- 4 parameters with t-statistic > 2
  - clogP
  - MolWeight
  - RotBonds
  - AromProp

$$\text{Log}(S_w) = 0.16 - 0.63 \times \text{clogP} - 0.0062 \times \text{MolWeight} + 0.066 \times \text{RotBonds} - 0.74 \times \text{AromProp}$$

J. S. Delaney, “ESOL: Estimating Aqueous Solubility Directly from Molecular Structure,” *Journal of Chemical Information and Modeling*, vol. 44, no. 3, pp. 1000–1005, May 2004.

RDKit Implementation

RDKit Based Implementation 1/2

- Use Regression Equation:

$$\text{Log}(S_w) = 0.16 - 0.63 \times \text{clogP} - 0.0062 \times \text{MolWeight} + 0.066 \times \text{RotBonds} - 0.74 \times \text{AromProp}$$

- Calculate properties using RDKit.

University of Cyprus
Department of Computer Science

RDKit Based Implementation 2/2

```
"""
This module calculates the aqueous (water) solubility based on::

    J. S. Delaney, "ESOL: Estimating Aqueous Solubility Directly from
    Molecular Structure," Journal of Chemical Information and Modeling,
    vol. 44, no. 3, pp. 1000-1005, May 2004.

The equation proposed by Delaney is::

    clogSw = 0.16
             - 0.63 x clogP
             - 0.0062 x MolWeight
             + 0.066 x RotBonds
             - 0.74 x AromProp

    clogP
        Octanol - water partition coefficient, lipophilicity factor.

    MolWeight
        Molecular Weight

    RotBonds
        Rotatable Bonds

    AromProp
        Aromatic Proportion, the proportion of heavy atoms (MolWeight > 1)
        in the molecule that are in an aromatic ring.

@author: Christos Kannas
"""
```

RDKit Based Implementation 2/2

```
def clogSw(molObj):
    """
    Calculate an estimation of water solubility.

    Input: RDKit molecule object

    Output: clogSw estimation
    """
```

```
    return clogSw_value
```

University of Cyprus
Department of Computer Science

RDKit Based Implementation 2/2

```
aromaticSmarts = "[a]"
```

```
    from rdkit import Chem
    from rdkit.Chem import Descriptors

    # calculate MolWeight
    MolWeight = Descriptors.MolWt(molObj)
    # calculate clogP
    clogP = Descriptors.MolLogP(molObj)
    # calculate RotBonds
    RotBonds = Descriptors.NumRotatableBonds(molObj)
    # calculate the number of aromatic heavyatoms in the molecule
    aromaticHeavyatoms = len(molObj.GetSubstructMatches(
                            Chem.MolFromSmarts(aromaticSmarts)))
    # calculate total number of atoms in the molecule
    numAtoms = molObj.GetNumAtoms()
    # calculate Aromatic Proportion
    AromProp = float(aromaticHeavyatoms) / numAtoms

    # then calculate clogSw...
    clogSw_value = 0.16 \
                   - 0.63 * clogP \
                   - 0.0062 * MolWeight \
                   + 0.066 * RotBonds \
                   - 0.74 * AromProp
```

Results

University of Cyprus
Department of Computer Science

RDKit
Open-Source Cheminformatics and Machine Learning

SEVENTH FRAMEWORK PROGRAMME

LiSIs
Life Sciences Informatics System

Testing…

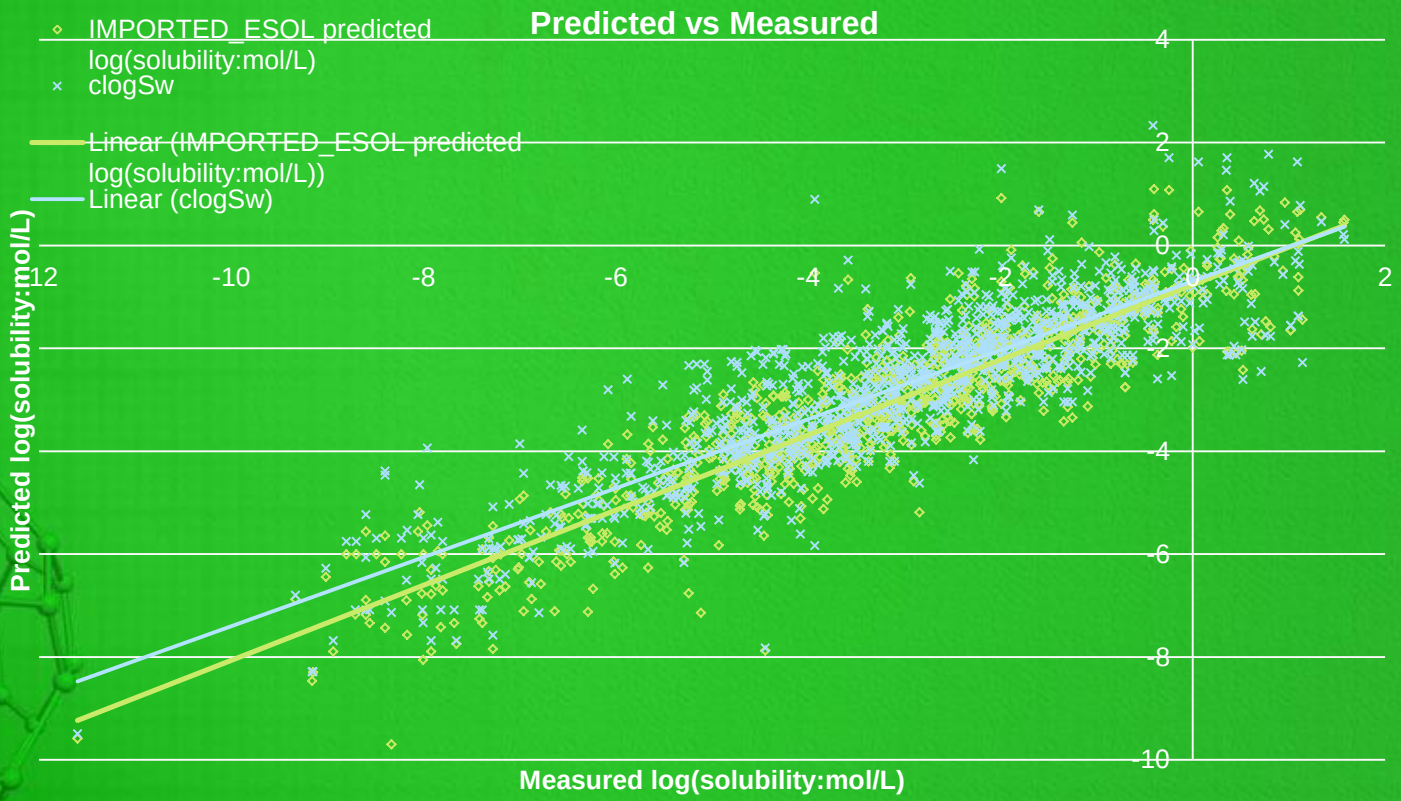
- Supplementary Dataset:
  - 1143 molecules with:
    - Measured Water Solubility (logSw)
    - ESOL

- Correlation Charts:
  - Measured vs ESOL
  - Measured vs RDKit_clogSw
  - ESOL vs RDKit_clogSw
  - Measured vs ESOL vs RDKit_clogSw

University of Cyprus
Department of Computer Science

Correlation Table & Chart

	IMPORTED_measured log(solubility:mol/L)	IMPORTED_ESOL predicted log(solubility:mol/L)	clogSw
IMPORTED_measured log(solubility:mol/L)	1		
IMPORTED_ESOL predicted log(solubility:mol/L)	0.90794375	1	
clogSw	0.864718601	0.964683313	1



University of Cyprus
Department of Computer Science

RDKit
Open-Source Cheminformatics and Machine Learning

SEVENTH FRAMEWORK PROGRAMME

LiSIs
Life Sciences Informatics System

Conclusion

- Comparable results.
- Easy, fast and relatively accurate.
- What is importance of adding Hydrogens prior to Aromatic Proportion calculation?