The Kaggle EUOS/SLAS Solubility Challenge: Visualizing and Understanding The Data Helps in Modelling

Bernhard Rohde

AIDD School, Helsinki, March 23, 2023

Overview

- Who am I?
- The EUOS/SLAS Solubility Challenge
- Instance Based Offset Learning
- Application of IBOL to Kaggle Challenge
- Analyzing Dataset
- Competition Entry: Prior-Only IBOL
- Post-Deadline Results
- Lessons Learned

Who am I?

- Diploma thesis on simulating EPR spectra in frozen Argon (1981, G. Maier/Gießen/DE, 'The allyl radical **is** flat')
- PhD on canonicalizing and searching of chemical structure representations (1982-87, A.S. Dreiding/Zürich/CH)
- Working for Ciba-Geigy => Ciba => Novartis (4/1987 to 2/2021), now part-time consultant
- Projects:
  - Computer Assisted Synthesis Planning (CASP, '87-'91, Poor Man's Synthesis Planning)
  - Structure Registration (CESAR/MACCS, CERES/ISIS, WITCH/Custom, SMR/Custom, CICLOPS/Custom Combichem)
  - Med Chem Databases (Delphi/Custom, WinMerlin/Daylight, Avalon/Oracle+Custom Cartridge, CDF-DART/Oracle+ChemAxon Cartridge)
- Open-Source Tools:
  - Avalon Toolkit
    - STRUCHK ('88): Structure checking, and standardization
    - Depicter: Used for WinMerlin, Avalon, DART, Web Service
  - Avalon Tools in RDKit: Fingerprinting, Canonicalization
- Research Interests:
  - Mostly Bayesian Methods, but trying to recycle/re-apply the above skills
  - Bayesian Optimization of chemical structures for docking (with Morgan Thomas)
  - Probabilistic Lead Optimization Flowchart (multi-objective BO, retrospective)

The EUOS/SLAS Solubility Challenge

- The Data
  - Nephelometric classification based on control compounds (Amiodarone, Phenytoin) at 10 µM in PBS (pH 7.4)
  - ~70'000 compound classifications for training, ~30'000 compounds for test and ranking
  - 352 compounds per plate (2x16 (edge?) positions used for positive and negative controls)
  - Screens run in duplicate with identical position on plate. Classification based on average.
  - Post challenge information: Compounds in 'low' category were confirmed separately and only the confirmation result had been reported in the dataset.
- The Rules
  - Quadratically weighted kappa coefficient
  - Preliminary ranking by ('public') 50% of test data
  - Final ranking by other ('private') 50% of test data
  - Implied constraints
    - No information beyond challenge dataset
    - Desired solution should be based on structural information

Sketch of method data

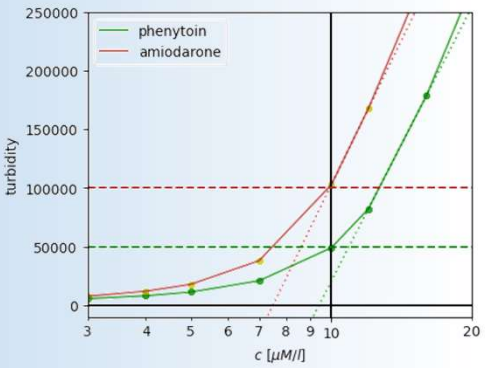
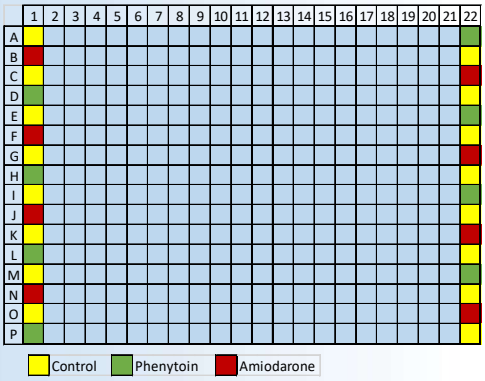


Figure 1 in Analytical Chemistry, Vol. 81, No. 8, April 15, 2009 shows turbidity signal for different concentrations of single compound.

The EUOS/SLAS Solubility Challenge

- The Data
  - Nephelometric classification based on control compounds (Amiodarone, Phenytoin) at 10 µM in PBS (pH 7.4)
  - ~70'000 compound classifications for training, ~30'000 compounds for test and ranking
  - 352 compounds per plate (2x16 (edge?) positions used for positive and negative controls)
  - Screens run in duplicate with identical position on plate. Classification based on average.
  - Post challenge information: Compounds in 'low' category were confirmed separately and only the confirmation result had been reported in the dataset.
- The Rules
  - Quadratically weighted kappa coefficient
  - Preliminary ranking by ('public') 50% of test data
  - Final ranking by other ('private') 50% of test data
  - Implied constraints
    - No information beyond challenge dataset
    - Desired solution should be based on structural information

Example Plate Layout



DMSO-Controls – Phenytoin: high (2)
Phenytoin – Amiodarone: medium (1)
Amiodarone - oo: low (0)
Acoustic dispenser: Echo550 https://www.selectscience.net/SelectScience-TV/Videos/echo-liquid-handling-systems-demonstration/?videoID=149

The EUOS/SLAS Solubility Challenge

- The Data
  - Nephelometric classification based on control compounds (Amiodarone, Phenytoin) at 10 µM in PBS (pH 7.4)
  - ~70'000 compound classifications for training, ~30'000 compounds for test and ranking
  - 352 compounds per plate (2x16 (edge?) positions used for positive and negative controls)
  - Screens run in duplicate with identical position on plate. Classification based on average.
  - Post challenge information: Compounds in 'low' category were confirmed separately and only the confirmation result had been reported in the dataset.
- The Rules
  - Quadratically weighted kappa coefficient
  - Preliminary ranking by ('public') 50% of test data
  - Final ranking by other ('private') 50% of test data
  - Implied constraints
    - No information beyond challenge dataset
    - Desired solution should be based on structural information

$$W_{i,j} = \begin{bmatrix} 0 & 0.25 & 1 \\ 0.25 & 0 & 0.25 \\ 1 & 0.25 & 0 \end{bmatrix}$$

$C_{i,j}$: $Confusion\ Matrix$

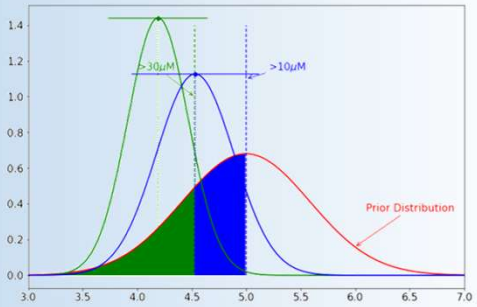
$E_{i,j}$: $Expected\ Confusion\ Matrix\ for\ Random\ Class\ Assignment$

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j} C_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}$$

DMSO-Controls – Phenytoin: high (2)
Phenytoin – Amiodarone: medium (1)
Amiodarone - oo: low (0)
Acoustic dispenser: Echo550 https://www.selectscience.net/SelectScience-TV/Videos/echo-liquid-handling-systems-demonstration/?videoID=149

Instance-Based Offset Learning

- "Simple" yard-stick model for realistic medicinal chemistry data
  - Unbalanced, lots of inactive compounds
  - **Censored data e. g. '> 10 µM', measurements with error**
  - "Switch to classification" is wrong reflex
  - Provides error estimates to map applicability domain
- K-Nearest Neighbor Regression on Bayesian Steroids
- Relevance Kernel ρ(sim(s,s_n))
- Powerset mixture model with KL-optimal Gaussian prediction
- Neighbor-derived mixture components can be combined in various ways, e. g. as even mixtures or consensus of experts.
- Regression Model of Prior Mean
- Regression Model of Neighbor Offsets
- (Initially) finite difference gradients for Maximum Likelihood optimization
- Model likelihood for selection of regressors and fingerprint generators
- Data likelihood using uncertain and censored data points
- Bayesian Information Criterion to regularize parameter optimization
- Greedily optimize BIC by adding and removing parameters and FP generators



Instance-Based Offset Learning

- "Simple" yard-stick model for realistic medicinal chemistry data
  - Unbalanced, lots of inactive compounds
  - Censored data e. g. '> 10 µM', measurements with error
  - "Switch to classification" is wrong reflex for censored data
  - Provides error estimates to map applicability domain
- **K-Nearest Neighbor Regression on Bayesian Steroids**
- **Relevance Kernel ρ(sim(s,s_n))**
- **Powerset mixture model with KL-optimal Gaussian prediction**
- Neighbor-derived mixture components can be combined in various ways, e. g. as even mixtures or consensus of experts.
- Regression Model of Prior Mean
- Regression Model of Neighbor Offsets
- (Initially) finite difference gradients for Maximum Likelihood optimization
- Model likelihood for selection of regressors and fingerprint generators
- Data likelihood using uncertain and censored data points
- Bayesian Information Criterion to regularize parameter optimization
- Greedily optimize BIC by adding and removing parameters and FP generators

$$\mathfrak{D} = \{d_i = (q_i, \mu_i, \sigma_i)\}_{i=1..N}$$

$$\mathfrak{N}_K(s|\mathfrak{D}) = K \text{ nearest neighbors to } s \text{ from } \mathfrak{D}$$

$$P(y|s,\mathfrak{D}) = \sum_{\mathfrak{n}\subseteq\mathfrak{N}_K(s|\mathfrak{D})} \pi(\mathfrak{n}|s)P(y|\mathfrak{n})$$

$$\pi(\mathfrak{n}|s) = \prod_{n\in\mathfrak{n}} \rho(sim(s,s_n)) \prod_{n\in\mathfrak{N}_K(s|\mathfrak{D})\setminus\mathfrak{n}} (1-\rho(sim(s,s_n)))$$

$$P_{KL}(y|s) = N(y;\mu,\sigma)$$

$$\mu = \mathbb{E}_{P(y|s,\mathfrak{D})}(X)$$

$$\sigma = \sqrt{\mathbb{E}_{P(y|s,\mathfrak{D})}(X^2) - \mu^2}$$

Instance-Based Offset Learning

- "Simple" yard-stick model for realistic medicinal chemistry data
  - Unbalanced, lots of inactive compounds
  - Censored data e. g. '> 10 μM', measurements with error
  - "Switch to classification" is wrong reflex for censored data
  - Provides error estimates to map applicability domain
- K-Nearest Neighbor Regression on Bayesian Steroids
- Relevance Kernel ρ(sim(s,s_n))
- Powerset mixture model with KL-optimal Gaussian prediction
- Neighbor-derived mixture components can be combined in various ways, e. g. as **even mixtures** or **consensus of experts**.
- Regression Model of Prior Mean
- Regression Model of Neighbor Offsets
- (Initially) finite difference gradients for Maximum Likelihood optimization
- Model likelihood for selection of regressors and fingerprint generators
- Data likelihood using uncertain and censored data points
- Bayesian Information Criterion to regularize parameter optimization
- Greedily optimize BIC by adding and removing parameters and FP generators

$$P(y|\mathbf{n}) = |\mathbf{n}|^{-1} \sum_{n \in \mathbf{n}} N(y; \mu'_n, \sigma_n) \qquad \text{if } \mathbf{n} \neq \varnothing$$
$$= N(y; \mu'_p, \sigma_p) \qquad \text{if } \mathbf{n} = \varnothing$$

or

$$P(y|\mathbf{n}) = N(y; \mu_c, \sigma_c)$$
$$\frac{1}{\sigma_c^2} = \frac{1}{\sigma_p^2} + \sum_{n \in \mathbf{n}} \frac{1}{\sigma_p^2}$$
$$\mu_c = \sigma_c^2 \left(\frac{\mu'_p}{\sigma_p^2} + \sum_{n \in \mathbf{n}} \frac{\mu'_n}{\sigma_n^2} \right)$$

Instance-Based Offset Learning

- "Simple" yard-stick model for realistic medicinal chemistry data
  - Unbalanced, lots of inactive compounds
  - Censored data e. g. '> 10 µM', measurements with error
  - "Switch to classification" is wrong reflex for censored data
  - Provides error estimates to map applicability domain
- K-Nearest Neighbor Regression on Bayesian Steroids
- Relevance Kernel ρ(sim(s,s_n))
- Powerset mixture model with KL-optimal Gaussian prediction
- Neighbor-derived mixture components can be combined in various ways, e. g. as even mixtures or consensus of experts.
- **Regression Model of Prior Mean**
- **Regression Model of Neighbor Offsets**
- (Initially) finite difference gradients for Maximum Likelihood optimization
- Model likelihood for selection of regressors and fingerprint generators
- Data likelihood using uncertain and censored data points
- Bayesian Information Criterion to regularize parameter optimization
- Greedily optimize BIC by adding and removing parameters and FP generators

$$\mu'_n = \mu_n + (\boldsymbol{x}(s) - \boldsymbol{x}_n) \cdot \boldsymbol{f}$$
$$\mu'_p = \mu_p + (\boldsymbol{x}(s) - \bar{\boldsymbol{x}}) \cdot \boldsymbol{f}_p$$

Instance-Based Offset Learning

- "Simple" yard-stick model for realistic medicinal chemistry data
  - Unbalanced, lots of inactive compounds
  - Censored data e. g. '> 10 μM', measurements with error
  - "Switch to classification" is wrong reflex for censored data
  - Provides error estimates to map applicability domain
- K-Nearest Neighbor Regression on Bayesian Steroids
- Relevance Kernel ρ(sim(s,s_n))
- Powerset mixture model with KL-optimal Gaussian prediction
- Neighbor-derived mixture components can be combined in various ways, e. g. as even mixtures or consensus of experts.
- Regression Model of Prior Mean
- Regression Model of Neighbor Offsets
- (Initially) finite difference gradients for Maximum Likelihood optimization
- **Model likelihood for selection of regressors and fingerprint generators**
- Data likelihood using uncertain and censored data points
- Bayesian Information Criterion to regularize parameter optimization
- Greedily optimize BIC by adding and removing parameters and FP generators

$$P(M|f_p, f, g) = \frac{1-\lambda_{f_p}}{1-\lambda_{f_p}^{N_f}} \lambda_{f_p}^{M_{f_p}} \binom{M_{f_p}}{N_f}^{-1} \times \frac{1-\lambda_f}{1-\lambda_f^{N_f}} \lambda_f^{M_f} \binom{M_f}{N_f}^{-1} \times \frac{1-\lambda_g}{1-\lambda_g^{N_g}} \lambda_g^{M_g} \binom{M_g}{N_g}^{-1}$$

Instance-Based Offset Learning

- "Simple" yard-stick model for realistic medicinal chemistry data
  - Unbalanced, lots of inactive compounds
  - Censored data e. g. '> 10 μM', measurements with error
  - "Switch to classification" is wrong reflex for censored data
  - Provides error estimates to map applicability domain
- K-Nearest Neighbor Regression on Bayesian Steroids
- Relevance Kernel ρ(sim(s,s_n))
- Powerset mixture model with KL-optimal Gaussian prediction
- Neighbor-derived mixture components can be combined in various ways, e. g. as even mixtures or consensus of experts.
- Regression Model of Prior Mean
- Regression Model of Neighbor Offsets
- (Initially) finite difference gradients for Maximum Likelihood optimization
- Model likelihood for selection of regressors and fingerprint generators
- **Data likelihood using uncertain and censored data points**
- **Bayesian Information Criterion to regularize parameter optimization**
- **Greedily optimize BIC by adding and removing parameters and FP generators**

$$\mathfrak{D}_n = \{(q_i, \mu_i, \sigma_i) \in \mathfrak{D} | i < n\}$$

$$L(\mathfrak{D}) = \prod_{d_n | q_n =' <'} \int_{-\infty}^{\mu_n} P_{KL}(y|s, \mathfrak{D}_n) dy \times \prod_{d_n | q_n =' >'} \int_{\mu_i}^{\infty} P_{KL}(y|s, \mathfrak{D}_n) dy \times \prod_{d_n | q_n =' ='} \int_{-\infty}^{\infty} N(y' - y; \mu_n, \sigma_n)\, P_{KL}(y|s, \mathfrak{D}_n) dy$$

$$\hat{L}(\mathfrak{D}) = \max_{\mu_\rho, \sigma_\rho, f_\rho, f} L(\mathfrak{D})$$

$$BIC = M_f ln(|\mathfrak{D}|) - 2ln(\hat{L}(\mathfrak{D})) - 2ln(P(M|f_\rho, f, g))$$

Application of IBOL to Kaggle Challenge Structure Preprocessing

- RDKit
  - Canonical representation of functional groups and salts
  - Isolation of main fragment (assuming counter ions don't affect kinetic solubility in buffer)
  - Assign FG categories (Acids, Amines, Aromatics, Quaternary Ammonium)
  - Compute Cheminformatics descriptors (clogp, cmr, tpsa, nrb, maxpc, minpc, diameter, radius)
  - Avalon fingerprint calculation
  - Precompute near neighbor lists

Application of IBOL to Kaggle Challenge Naïve First Try

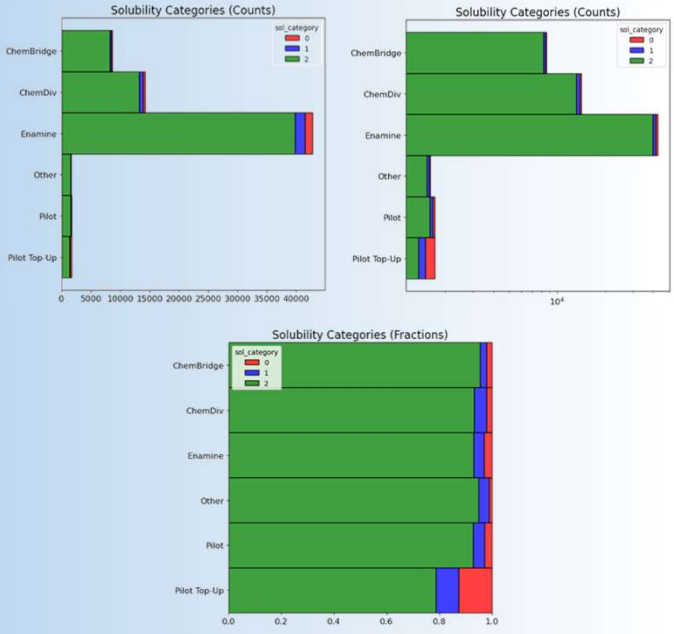
- Approach for challenge:
  - Extend censored data use to include ranges
  - Use log of class read-out limits as range boundaries
  - Use full fingerprints for similarity
  - Try CLOGP, CMR, and fCSP3 as single regressors
  - Use 20’000 training rows (for speed reasons)
  - Choose most likely predicted solubility class

Application of IBOL to Kaggle Challenge Naïve First Try

- Approach for challenge:
  - Extend censored data use to include ranges
  - Use log of class read-out limits as range boundaries
  - Use full fingerprints for similarity
  - Try CLOGP, CMR, and fCSP3 as single regressors
  - Use 20'000 training rows (for speed reasons)
  - Choose most likely predicted solubility class
- Result: $\kappa = 0.004$ ☹
  - Only classes 1 and 2 were populated
    => kappa optimization
  - There was supplier information on the compounds available to correct for supplier bias
    (https://www.eu-openscreen.eu/services/compound-collection.html)
    => $\kappa = 0.08561$
  - Competition was much better but not spectacular
    => analyze dataset and improve approach

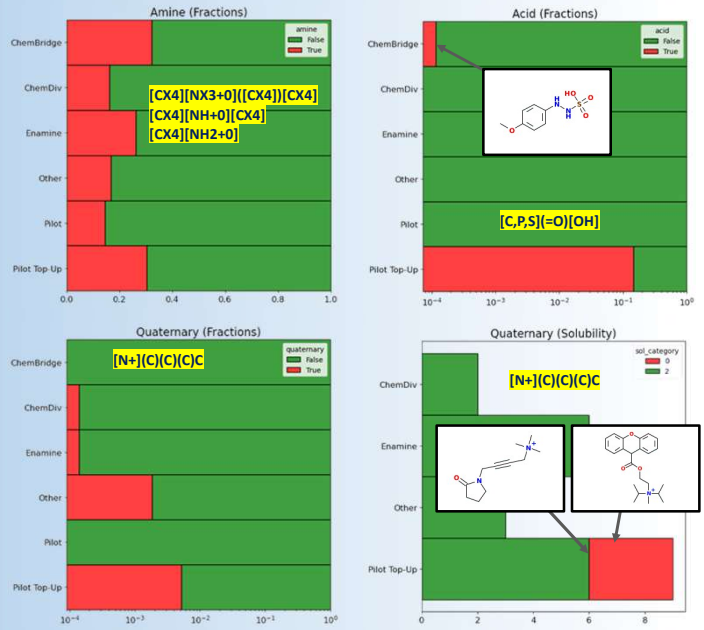
Analyzing Dataset

- Dataset by Supplier
  - Solubility classes
    - Most compounds from ‘Enamine’, ‘ChemDiv’, ‘ChemBridge’
    - Two different ‘Pilot’ sets
    - Category fractions unevenly distributed
    - Pilot Top-Up set has more informative members



Analyzing Dataset

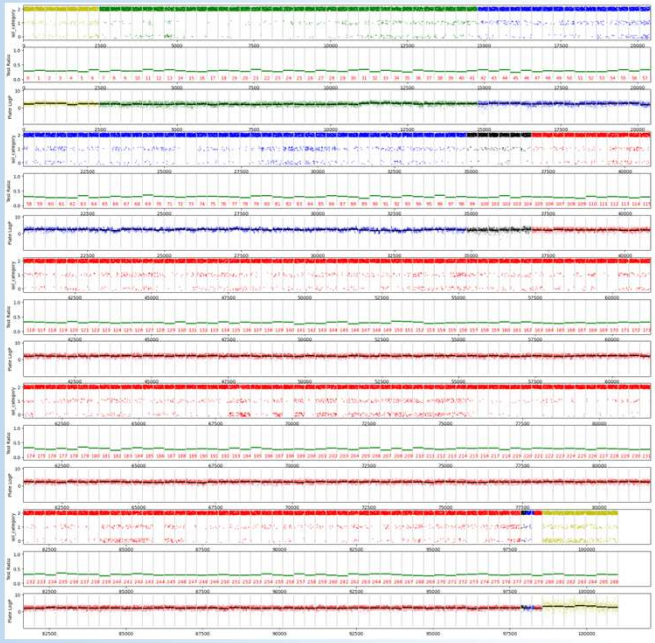
- Dataset by Supplier
  - Solubility classes
  - Functional groups
    - Amines reasonable well distributed
    - Acids (almost) exclusively in Top-Up set
    - Quaternary amines are rare, but some are ‘insoluble’ in Top-Up set, which is odd



https://www.tocris.com/products/oxotremorine-m_1067 Solubility > 100 mM
https://cdn.caymanchem.com/cdn/insert/23609.pdf ~22 mM

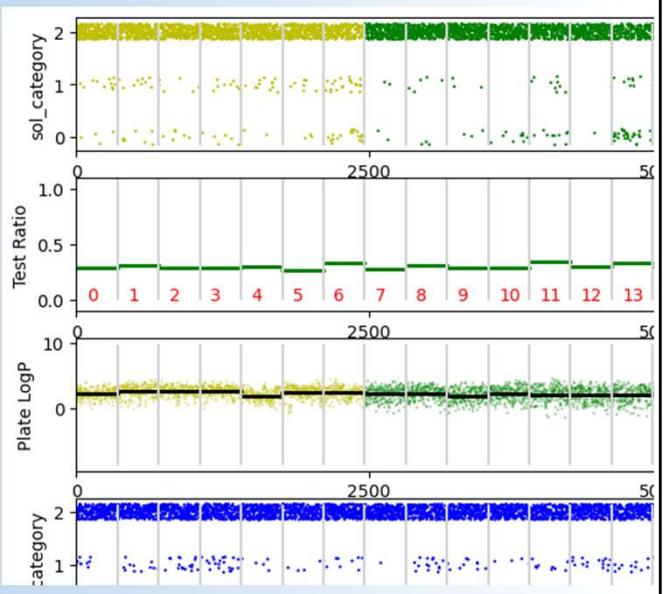
Analyzing Dataset

- Dataset by Plate
  - Guessing plate numbers
    - Consecutive EOS-Ids in two ranges:
      [EOS1-EOS98560]
      [EOS100001-EOS102459]
    - Plates contain
      352 compounds + 16 controls
    - Duplicates for averaging
      are plate copies



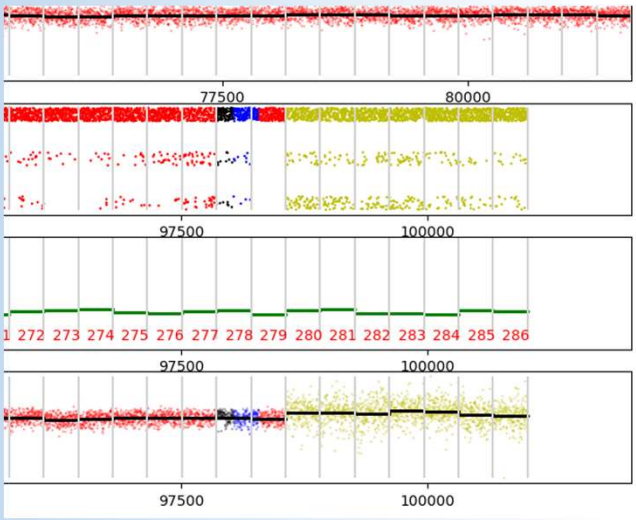
Analyzing Dataset

- Dataset by Plate
  - Guessing plate numbers
    - Consecutive EOS-Ids in two ranges:
      [EOS1-EOS98560]
      [EOS100001-EOS102459]
    - Plates contain
      352 compounds + 16 controls
    - Duplicates for averaging
      are plate copies



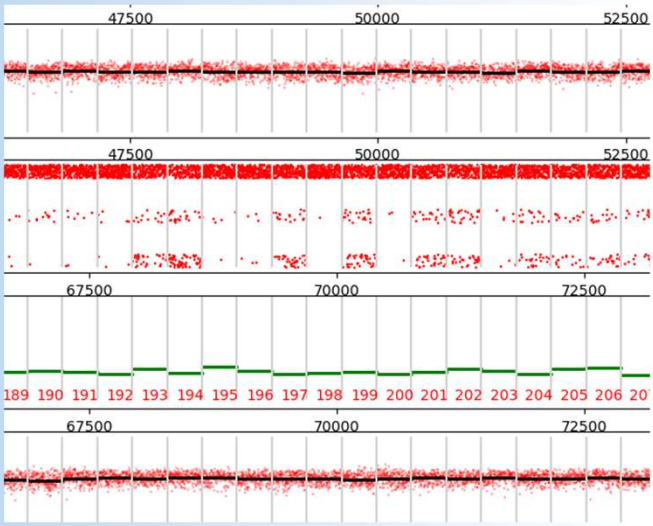
Analyzing Dataset

- Dataset by Plate
  - Guessing plate numbers
    - Consecutive EOS-Ids in two ranges:
      [EOS1-EOS98560]
      [EOS100001-EOS102459]
    - Plates contain
      352 compounds + 16 controls
    - Duplicates for averaging
      are plate copies



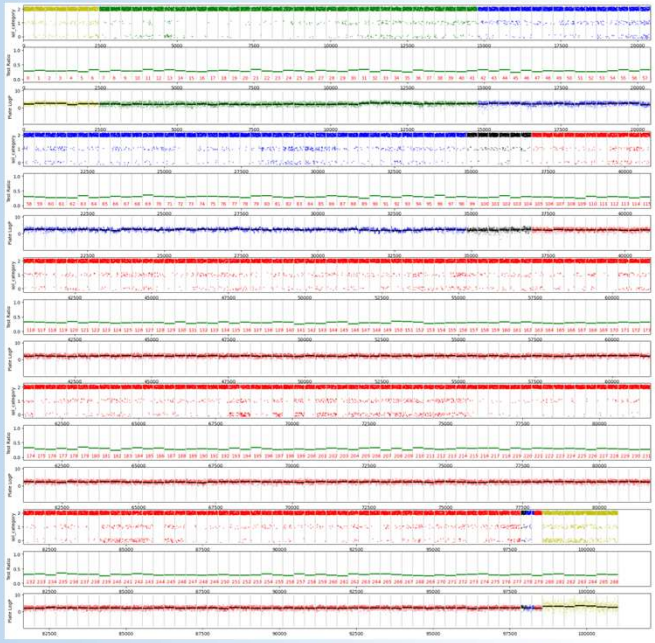
Analyzing Dataset

- Dataset by Plate
  - Guessing plate numbers
    - Consecutive EOS-Ids in two ranges:
      [EOS1-EOS98560]
      [EOS100001-EOS102459]
    - Plates contain
      352 compounds + 16 controls
    - Duplicates for averaging
      are plate copies



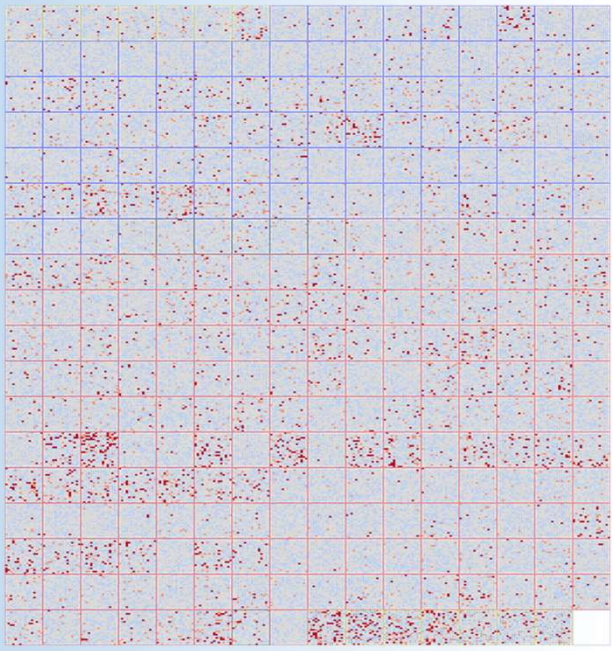
Analyzing Dataset

- Dataset by Plate
  - Guessing plate numbers
    - Consecutive EOS-Ids in two ranges:
      [EOS1-EOS98560]
      [EOS100001-EOS102459]
    - Plates contain
      352 compounds + 16 controls
    - Duplicates for averaging
      are plate copies



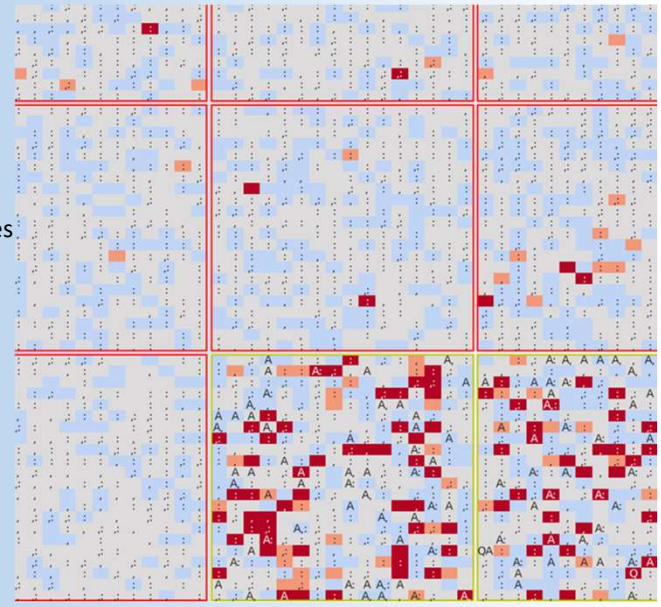
Analyzing Dataset

- Dataset by Plate
  - Guessing plate numbers
    - Consecutive EOS-Ids in two ranges:
      [EOS1-EOS98560]
      [EOS100001-EOS102459]
    - Plates contain
      352 compounds + 16 controls
    - Duplicates for averaging
      are plate copies



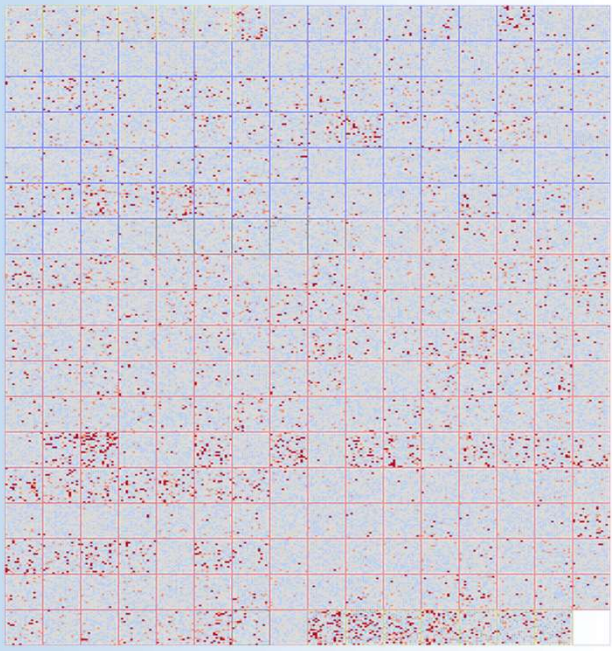
Analyzing Dataset

- Dataset by Plate
  - Guessing plate numbers
    - Consecutive EOS-Ids in two ranges
      [EOS1-EOS98560]
      [EOS100001-EOS102459]
    - Plates contain
      352 compounds + 16 controls
    - Duplicates for averaging
      are plate copies



Analyzing Dataset

- Dataset by Plate
  - Guessing plate numbers
    - Consecutive EOS-Ids in two ranges:
      [EOS1-EOS98560]
      [EOS100001-EOS102459]
    - Plates contain
      352 compounds + 16 controls
    - Duplicates for averaging
      are plate copies



Competition Entry: Prior-Only IBOL Model

- Just model plate, row, column, and clogp parameters
  - Gaussian prediction $N(x|\mu_i,\sigma)$
  - Common σ
  - Prior μ_0
  - 286 plate offsets $\delta\mu_{p(i)}$
  - 16 row offsets $\delta\mu_{p(i)}$
  - 22 column offsets $\delta\mu_{c(i)}$
  - Separate model for base and top-up data
  - f_{clogp} * $clogp_i$ (centered and normalized clogp)
  - All parameters regularized by normal prior
  - κ-optimized probability limits
  - Result: $\kappa_{public} = 0.19395$, $\kappa_{private} = 0.21748$

$$\mu_i = \mu_0 + \delta\mu_{p(i)} + \delta\mu_{r(i)} + \delta\mu_{c(i)} + f_{clogp} clogp_i$$

$$P(i \mid s_i = l) = \int_{v_l}^{\infty} N(v|\mu_i, \sigma)\, dv$$

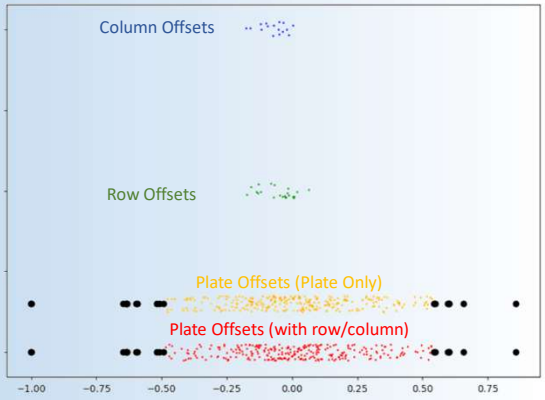
$$P(i \mid s_i = m) = \int_{v_m}^{v_l} N(v|\mu_i, \sigma)\, dv$$

$$P(i \mid s_i = h) = \int_{-\infty}^{v_m} N|\mu_i, \sigma)\, dv$$

$$logMAP = logP(\mu_0) + logP(\sigma) + \sum_p logP(\delta\mu_p) + \sum_r logP(\delta\mu_r) + \sum_c logP(\delta\mu_c) + logP(f_{clogp}) + \sum_{i|s_i=l} logP(i|s_i = l) + \sum_{i|s_i=m} logP(i|s_i = m) + \sum_{i|s_i=h} logP(i|s_i = h)$$

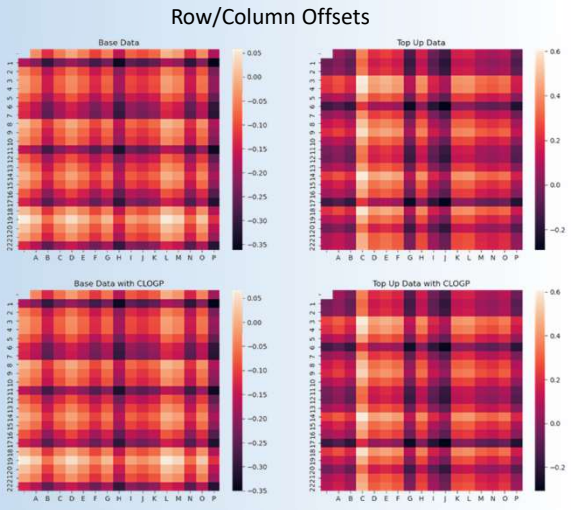
Competition Entry: Prior-Only IBOL Model

- Just model plate, row, column, and clogp parameters
  - Gaussian prediction $N(x|\mu_i,\sigma)$
  - Common σ
  - Prior μ_0
  - **286 plate offsets $\delta\mu_{p(i)}$**
  - **16 row offsets $\delta\mu_{p(i)}$**
  - **22 column offsets $\delta\mu_{c(i)}$**
  - Separate model for base and top-up data
  - f_{clogp} * $clogp_i$ (centered and normalized clogp)
  - All parameters regularized by normal prior
  - κ-optimized probability limits
  - Result: κ_{public} = 0.19395, $\kappa_{private}$ = 0.21748



Competition Entry : Prior-Only IBOL Model

- Just model plate, row, column, and clogp parameters
  - Gaussian prediction $N(x|\mu_i,\sigma)$
  - Common σ
  - Prior μ_0
  - 286 plate offsets $\delta\mu_{p(i)}$
  - **16 row offsets $\delta\mu_{p(i)}$**
  - **22 column offsets $\delta\mu_{c(i)}$**
  - Separate model for base and top-up data
  - $f_{clogp} * clogp_i$ (centered and normalized clogp)
  - All parameters regularized by normal prior
  - κ-optimized probability limits
  - Result: $\kappa_{public} = 0.19395$, $\kappa_{private} = 0.21748$



Competition Entry : Prior-Only IBOL Model

- Just model plate, row, column, and clogp parameters
  - Gaussian prediction $N(x|\mu_i,\sigma)$
  - Common σ
  - Prior μ_0
  - 286 plate offsets $\delta\mu_{p(i)}$
  - 16 row offsets $\delta\mu_{p(i)}$
  - 22 column offsets $\delta\mu_{c(i)}$
  - Separate model for base and top-up data
  - $f_{clogp} * clogp_i$ (centered and normalized clogp)
  - All parameters regularized by normal prior
  - **κ-optimized probability limits**
  - Result: $\kappa_{public} = 0.19395$, $\kappa_{private} = 0.21748$

$$c_{sol} = \begin{cases} 0, & \text{if } N(y > log_{10}(100000); \mu_i, \sigma_i) \geq 1 - P_{low}, else \\ 1, & \text{if } N(y > log_{10}(50000); \mu_i, \sigma_i) \geq 1 - P_{med}, \\ 2, & \text{otherwise.} \end{cases}$$

$$P_{low} = 0.847$$
$$P_{med} = 0.86459$$

Optimized by MH-like random search

Post Deadline Results

- Learned to use AutoGrad and JAX
  - AutoGrad provides an easy substitute for numpy
  - AutoGrad team moved to JAX
  - JAX is almost as easy as AutoGrad, but has some 'Sharp Bits'
  - Rewritten most of original IBOL tools to work on DataFrames with JAX
    => Plug-And-Play tool

https://jax.readthedocs.io/en/latest/notebooks/Common_Gotchas_in_JAX.html

Post Deadline Results

- Dependence on Regression Components
  - **Using CLOGP or CMR yield private score < 0.04**
  - The submission entry (private score 0.2189) can be improved by adding CMR and FCSP3 as regressors
  - Just optimizing plate offsets yields most of the modeling power.
  - Lower Limit and Upper Limit are probability cutoffs optimized for κ on the training data.

Model Components	MAP Score	Kappa Train	Lower Limit	Upper Limit	Public Score	Private Score
CLOGP	21277.22	0.04334	0.93043	0.93246	0.04192	0.02923
CMR	21258.17	0.04985	0.92308	0.92713	0.06399	0.03822
Plate	19608.21	0.21072	0.86857	0.87046	0.19448	0.19768
Plate, Row, Column	19475.75	0.22713	0.85684	0.86396	0.19252	0.21781
Plate, Row, Column, CLOGP	19441.63	0.23196	0.847	0.86459	0.19565	0.2189
Plate, Row, Column, CMR	19401.01	0.23478	0.85036	0.86205	0.20293	0.21502
Plate, Row, Column, CLOGP, CMR	19391.69	0.2376	0.84755	0.85799	0.20307	0.22241
Plate, Row, Column, CLOGP, CMR, FCSP3	19385.27	0.23704	0.84047	0.86171	0.20715	0.22434

Post Deadline Results

- Dependence on Regression Components
  - Using CLOGP or CMR yield private score < 0.03
  - **The submission entry (private score 0.2189) can be improved by adding CMR and FCSP3 as regressors**
  - Just optimizing plate offsets yields most of the modeling power.
  - Lower Limit and Upper Limit are probability cutoffs optimized for κ on the training data.

Model Components	MAP Score	Kappa Train	Lower Limit	Upper Limit	Public Score	Private Score
CLOGP	21277.22	0.04334	0.93043	0.93246	0.04192	0.02923
CMR	21258.17	0.04985	0.92308	0.92713	0.04192	0.02923
Plate	19608.21	0.21072	0.86857	0.87046	0.19448	0.19768
Plate, Row, Column	19475.75	0.22713	0.85684	0.86396	0.19252	0.21781
Plate, Row, Column, CLOGP	19441.63	0.23196	0.847	0.86459	0.19565	0.2189
Plate, Row, Column, CMR	19401.01	0.23478	0.85036	0.86205	0.20293	0.21502
Plate, Row, Column, CLOGP, CMR	19391.69	0.2376	0.84755	0.85799	0.20307	0.22241
Plate, Row, Column, CLOGP, CMR, FCSP3	19385.27	0.23704	0.84047	0.86171	0.20715	0.22434

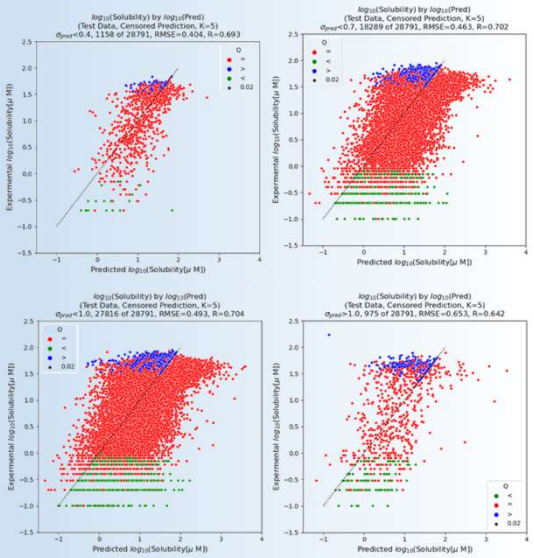
Post Deadline Results

- Dependence on Regression Components
  - Using CLOGP or CMR yield private score < 0.03
  - The submission entry (private score 0.2189) can be improved by adding CMR and FCSP3 as regressors
  - **Just optimizing plate offsets yields most of the modeling power.**
  - Lower Limit and Upper Limit are probability cutoffs optimized for κ on the training data.

Model Components	MAP Score	Kappa Train	Lower Limit	Upper Limit	Public Score	Private Score
CLOGP	21277.22	0.04334	0.93043	0.93246	0.04192	0.02923
CMR	21258.17	0.04985	0.92308	0.92713	0.04192	0.02923
Plate	19608.21	0.21072	0.86857	0.87046	0.19448	0.19768
Plate, Row, Column	19475.75	0.22713	0.85684	0.86396	0.19252	0.21781
Plate, Row, Column, CLOGP	19441.63	0.23196	0.847	0.86459	0.19565	0.2189
Plate, Row, Column, CMR	19401.01	0.23478	0.85036	0.86205	0.20293	0.21502
Plate, Row, Column, CLOGP, CMR	19391.69	0.2376	0.84755	0.85799	0.20307	0.22241
Plate, Row, Column, CLOGP, CMR, FCSP3	19385.27	0.23704	0.84047	0.86171	0.20715	0.22434

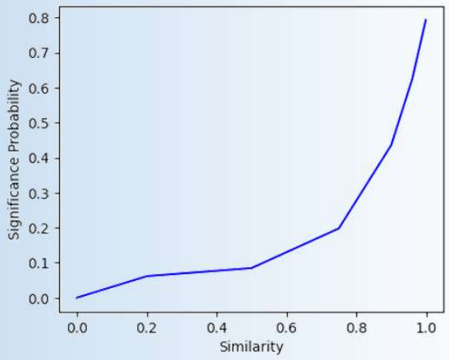
Post Deadline Results

- PubChem Solubility Assay AID1996
  - Filtered solute measured using chemiluminescent nitrogen detection
    => must contain nitrogen
  - 57.8 K compounds from
    Molecular Libraries Small Molecule Repository (NIH), Deposited Oct-2009
  - Random Split: 50% Training / 50% Test
  - IBOL Model with 5 Neighbors
  - Reimplemented using JAX and Dataframes
  - 14 Prior Regressors, ['FractionCSP3',…]
    13 Neighbor Regressors, ['MolLogP', …]
  - Optimized Avalon FP (9 of 18 generators)
    `ATOM_SYMBOL_PATH, AUGMENTED_ATOM, HCOUNT_PATH, HCOUNT_CLASS_PATH, HCOUNT_PAIR, RING_SIZE_COUNTS, FEATURE_PAIRS,SCAFFOLD_IDS, SCAFFOLD_COLORS`



Post Deadline Results

- PubChem Solubility Assay AID1996
  - Filtered solute measured using chemiluminescent nitrogen detection
    => must contain nitrogen
  - 57.8 K compounds from
    Molecular Libraries Small Molecule Repository (NIH), Deposited Oct-2009
  - Random Split: 50% Training / 50% Test
  - IBOL Model with 5 Neighbors
  - Reimplemented using JAX and Dataframes
  - 14 Prior Regressors, ['FractionCSP3',…]
    13 Neighbor Regressors, ['MolLogP', …]
  - Optimized Avalon FP (9 of 18 generators)
    `ATOM_SYMBOL_PATH, AUGMENTED_ATOM, HCOUNT_PATH, HCOUNT_CLASS_PATH, HCOUNT_PAIR, RING_SIZE_COUNTS, FEATURE_PAIRS,SCAFFOLD_IDS, SCAFFOLD_COLORS`



Lessons Learned

- Do not throw your favorite ML model at an arbitrary dataset and expect it to work.
- Having a competitor much ahead of you makes you think.
- Non-random draws from the structure universe can confuse modelling.
- Compare what you know about the problem with the data to spot (detrimental/exploitable) peculiarities. Visualization is key.
- As with many puzzles, there is more information than you think in the problem description.
- Finding the major sources of variance can give you a lot of mileage even if it does not help understanding the scientific problem.
- Probabilistic models can be used to predict uneven classification. The additional uncertainty model can even help in making decisions.
- Automatic differentiation is the key to learning (and “a retired dog can learn new tricks”).

Acknowledgements

- Challenge Organizers
  - Challenging dataset with lots to learning from
- Competitors
  - Interesting discussions
  - Providing the carrot for staying focused
- Morgan Thomas
  - Docking my IBOL-selected compounds for Bayesian Optimization
  - Discussing my odd intermediate results
- Igor Tetko
  - Invitation to give this presentation
- My wife
  - Living with a retiree who is busier than before retirement