SimulationsPlus

The Power of Data Sharing

Eric Jamois

11/14/2023

pKa Model Development

Problem Statement:

Public datasets for pKa data are limited to ca. 10,000 compounds w/ ca 14,000 pKa values

- Pharma companies have extensive amounts of data
  - ➔ Large potential for expansion of chemical space
  - ➔ Hurdles vs sensitivity of the data

But:

- Work through contract/legal associated w/ sharing of data
- Large amount of data curation typically required regardless of source
- Complex model construction process (high LOE): Quantum Leap

pKa Model Development with Bayer Pharma & Crop, Roche & Genentech

Started in 2012

JOURNAL OF CHEMICAL INFORMATION AND MODELING — Article — pubs.acs.org/jcim

Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve *in Silico* pK_a Prediction

Robert Fraczkiewicz,*[,†] Mario Lobell,*[,‡] Andreas H. Göller,[‡] Ursula Krenz,[‡] Rolf Schoenneis,[‡] Robert D. Clark,[†] and Alexander Hillisch[‡]

[†]Simulations Plus, Inc. 42505 10th Street West, Lancaster, California 93534, United States
[‡]Global Drug Discovery, Bayer Pharma AG, Wuppertal, Germany

Continued in 2022/2023

Simulations Plus Enters New Collaboration to Enhance Machine Learning Models for Ionization Constants (pKa)

August 04, 2022

Simulations Plus and Global Agrochemicals Leader to Collaborate on Machine Learning Models

Data sharing partnership will expand chemical coverage space and improve model performance in support of new approach methodologies to ensure product safety

February 16, 2023 08:30 AM Eastern Standard Time

New pKa Model Development

Solution Statement:

Joint development leading to:

- ➔ 14,176 ionization macroconstants from public sources
- ➔ 19,464 ionization macroconstants from Bayer alone

Additional data contributions from 2022/2023

- ➔ Roche: 19,000 compounds
- ➔ Genentech: 2,400 compounds
- ➔ Bayer Crop Sci: 4,100 compounds

RMSE Improvements:

0.792 ➔ 0.415

0.784 ➔ 0.519

1.146 ➔ 0.556

➔ **Total of 70,810 ionization constants**

Model Development Timeline

Timeline:

- Contracts Signed: 7/15/2022 and 10/3/2022
- Data Obtained: 7/20/2022 and later
- Model development: 7/20/2022 – 4/13/2023
- Final Model Selection: 4/14/2023
- Commercial Release: Jun. 2023

Business Model

- Data: Ownership to customer – License to SLP
- Model development: SLP
- Model validation & Publication: Joint

- Model(s): Ownership to SLP – License to customer
- Cost to customer: $0 | Royalty to customer: $0

Areas of Active  Exploration

ChemInformatics
- Tautomer-independent models
- Automated data curation
- AI technologies

Chemistry space
- BRo5 (PROTAC)
- Macrocyles

Models
- Clearance / microsomal stability
- Solubility
- Plasma Protein Binding
- Transporters
- CYP Induction
- Intestinal permeability
- UGT metabolism