Understanding Toxicokinetics Through Building A Jupyter Notebook

Jacob J. Ehrhardt

Department of Chemistry & Chemical Biology

CHEM 4PB3: Computational Models for Electronic Structure and Chemical Bonding

Dr. Rodrigo A. Vargas-Hernández

April 25, 2024

Section I: Introduction

Known as the central science, the study of chemistry strongly integrates concepts from both the physical and life sciences to understand processes that govern the physical world.[1] Data science is also a multidisciplinary field that derives value from data through the application of technological tools and scientific knowledge.[2] Recently, chemists have recognized the increasing value of programming in order to handle larger amounts of data, and apply new knowledge to make processes more efficient and sustainable.[3] In Computational Models for Electronic Structure and Chemical Bonding (CHEM 4PB3), tools for computational chemistry are interactively taught with the objective of making students more familiar with utilizing modern programming software packages to understand chemistry.[4] In Sustainable Chemistry: Analysis and Regulation (CHEM 4SC3); students are encouraged to develop the necessary skills to utilize numerical metrics related to sustainability and toxicology, then to understand and apply results by communicating findings.[5] The field of Environment Health and Safety (EHS) intersects these courses, due to the challenge of managing increasing chemical emissions, understanding novel chemical's acute and chronic toxicology, and applying this toxicological knowledge in the development of health and safety regulations.[5,6] In this context, programming and machine learning can be utilized in not just data science and analytics, but in the development of interpretable machine learning (IML) tools to employ computational toxicology, rapidly enhancing the efficiency, effectiveness, and safety of current practices.[7]

Toxicokinetics is the study of a toxic chemical's fate throughout the body from the point of intake all the way to its elimination from the body.[8] Toxicokinetics can neatly be broken down into four components; the absorption of the chemical into the body, its distribution throughout the body, metabolism or biotransformation of the initial chemical into a metabolite, and finally the elimination of the chemical through various excretory organs.[8–10] PROduction-To-Exposure

high-throughput (PROTEX-HT) is a large model that accurately simulates the toxicokinetic behaviours of various chemicals.[10] Given just the molecular structure and amount of chemical, PROTEX-HT is able to integrate simulated chemical transport in the environment and toxicokinetics in humans to calculate exposure and risk metrics for a chemical in various environments.[10]

Python is a high-level yet interpreted programming language, that offers ease of learning through helpful syntax, yet is still highly powerful.[2,3]. Its wide adoption has inspired the development of packages in many fields to aimed to enhance productivity and effectiveness. A Jupyter notebook is an interactive document that allows for the utilization of both code and text; making it an excellent tool for documenting explorative projects.[2]

Herein, a project utilizing a toxicokinetic dataset is integrated with a computational chemistry project aimed at developing python programming skills. Using Python in a Jupyter notebook, data analysis is able to be conducted and documented concurrently, allowing for a unique exploratory project that relays results like an interactive visual story.

Section II: Hypothesis

Through the gradual construction of a Jupyter Notebook, Python programming skills can be learned by performing data analysis on a PROTEX generated dataset. Jupyter's dual programming and documentation style allows for the learning of both programming skills and toxicokinetic principles. Through project-based learning, querying the dataset will allow for the natural discovery of the underlying toxicological principles, while concurrently extending upon these findings in CHEM 4SC3. Statistical analysis of the PROTEX dataset will yield insight into how demographic characteristics like age vary its toxicokinetic behaviour in humans. These insights can extend to include the influence of physiochemical properties on toxicokinetics, by evaluating chemical-dependent properties such as a chemical's molecular weight and K_{ow}.

Section III: Method Description

Learning commenced by creating a Jupyter Notebook in order to analyze the dataset and document findings within the same file. The notebook utilized linked titles organized by size using markdown formatting, creating structure between components and an interactive table of contents to navigate through sections more easily.

To begin, necessary libraries are pip installed under the title, allowing new readers to read the introductory text while the libraries are installing on their device. After describing, loading, and displaying the dataset using pandas, data analysis is conducted.

After inspecting each of the columns it was made clearer that the data described a simulated environment which measured the absorption, distribution, and elimination of 40 chemicals in humans aged 3, 14, and 25. Once these toxicokinetic processes were identified and understood, variables like demographic characteristics and chemical properties were evaluated for importance in each of these processes.

For absorption, the intake rate of the chemicals was compared against all three ages in order to determine if age had an influence on the amount of chemical a person consumes and through which route. Also, intake through all seven routes was summed and compared to the total chemical absorption value for each age in search of relevance.

For distribution, each chemical's concentrations in 4 bodily phases (lipids, blood, plasma, urine) were measured. The results of these, as well as the phases relationship to each other was analyzed to glean insight and information of the relevance of various properties' impact

For excretion, the chemicals with the highest concentrations in each of the distributed phases were identified, and their elimination through five different routes (egestion, metabolism, respiration, percutaneous, renal) were compared.

Section IV: Results

Using pandas, the chemical independent column titles are grouped according to their toxicokinetic process in Table 1.

Table 1. Chemical-independent values in the PROTEX dataset sorted under its appropriate ADME category.

Absorption	Chemical intake rates (ug/day)	Inhalation of outdoor air
		Inhalation of indoor air
		Dermal permeation
		Dietary ingestion
		Drinking water
		Ingestion of soil particles
		Mouthing ingestion
	Total chemical absorption (ug/day)	
Distribution	Lipid-normalized concentration (ng-chem/g-lipid)	
	Blood concentration (µM)	
	Plasma concentration (µM)	
	Urinary concentration (µM)	
Excretion	Steady-state total elimination half-life (hour)	
	Relative importance of elimination process (%)	Egestion
		Biotransformation (metabolism)
		Respiration
		Percutaneous
		Renal

As the sole value relating to metabolism in the given dataset, we listed biotransformation under the domain of excretion. In this context, biotransformation is considered a type of elimination because the concentration of the ingested chemical is decreasing as it is converted to its metabolite. In the case of volatile organic compounds (VOCs) for example, the metabolite is likely to be more water soluble, and therefore able to be eliminated through urine quicker than if it were stored as fat.[11] Since excretion modulates a chemical's toxic potential by eliminating it from the body, this reasoning can be applied to biotransformation.

For absorption, boxplots were created to visually compare the intake rates of all three ages for each of the seven intake routes. These plots made outlier chemicals for each pathway very obvious and helped to display a general visual trend for all pathways (slight increase in

intake rate with increasing age). However, the trend for mouthing ingestion was the opposite, with a large intake rate in 3 year old's followed by a steep drop-off in older ages. This warranted statistical investigation, so an ANOVA test was performed to determine if these differences between mean intake rates across different age groups was statistically significant. A code using pandas and SciPy outputted an easy to read chart specifying if the relationship was significant, and an automated text response to help the user understand the result. Therefore it was revealed that age has an effect on toxicokinetics because three year old's consume statistically significantly higher amounts of chemicals through mouthing ingestion than teenagers and adults.

It was also revealed that the sum of daily chemical intake though all seven pathways is greater than total chemical absorbance. This means that although a chemical may be intaken into the body, not all of it is absorbed, and therefore may not reach systemic circulation (carried through the circulatory system) and be available for distribution. However, the total chemical intake and absorption were highly correlated (R^2=0.999, code), and regression lines for both total intake and total absorption vs age were negatively sloped, meaning that older individuals intake and absorb fewer micrograms of chemicals per day.

For distribution, histograms were created for each of the four phases, comparing chemical concentration in each phase between the three ages. From the 12 histograms it was easy to tell that age did not change the distribution of the chemicals. In all histograms, majority (>60%) of the frequencies were in the lowest bin (0 to very little concentration in that phase), however the difference lies in the distribution of the remaining bins. For lipid, blood, and plasma concentrations, the remaining bins contained similar frequencies (<3) relatively evenly distributed across the x-axis. Urinary concentration was more binary, with chemicals either existing in the lowest two bins, or the lone bin at the far end of the x-axis (high concentration).

More insight into distribution was gleaned when the phases were compared against each other. Concentration of chemicals in lipids was compared against concentration in blood, plasma, and urine. From this we evaluated the strongest correlation between lipid and blood (R^2=0.91), a weaker correlation with plasma (R^2=0.75), and virtually no correlation with urine (R^2=0.01). This makes more sense when considering the water concentration in each phase; bodyfat is highly hydrophobic but still contains ~10-30% water[12], red blood cells in blood equate to ~52%,[13] plasma has 91-92%,[14] and urine is approximately 95% water[15]. For lipids, blood, and plasma, it is apparent that the relationship strengthens as the phase becomes less hydrophilic. However, for this to be true for urine a strong negative correlation should be present, but instead there is virtually no relationship. Combined with the binary distribution identified in the urine histograms, it seems as though a chemical's fate in urine is not governed by the same set of principles responsible for a chemicals fate in lipids, blood, and plasma.

In order to identify specific chemical properties involved, the chemicals with the highest concentration in each distribution phase were queried from the dataset, and using the CAS numbers provided in the column beside the chemical name; molecular weight, LogP, and canonical smiles were pulled into a chart. From this we can see that larger, more hydrophobic, halogenated chemicals reside in more hydrophobic phases.

The chemicals with the highest concentration in lipids (BDE 17), blood and plasma (PCDD 70) and urine (Cyanuric acid) had their excretion profiles compared. As age increased, all three chemicals saw egestion importance decrease while biotransformation increased. This can be explained due to the maturation of the liver as humans reach early adulthood, which allows for efficient metabolism of chemicals.[16] When considering lipophilicity, both BDE 17 and PCDD 70 with lipophilic LogP's (8.3 and 7.4 respectively) held majority of excretion importance with egestion, while the hydrophilic (LogP = -1.2) Cyanuric acid was excreted primarily through urine.

Section V: Summary

The explorative nature in creating the Jupyter Notebook enhanced python skills with each new code block, while the resulting analysis of the data yielded insight into the variables that do and do not impact toxicokinetics.

In terms of demographic characteristics, age was revealed to influence chemical intake, with young children ingesting statistically significantly higher levels through mouthing ingestion than adults, likely due to their curious nature and mouthing tendencies. It was also revealed that while a chemical's total intake and absorption are correlated, not all of an ingested chemical is systemically absorbed and able to be distributed throughout the body.

Chemical properties such as hydrophilicity/phobicity were also proven to influence a chemical's fate; with chemicals more averse to aqueous media consistently displaying higher concentrations in phases that lack water and have lipophilic properties. Modeling excretion displayed proof that independent of the chemical, egestion is downregulated as biotransformation matures with age from childhood to adulthood.

Combining multiple components, while PCDD 70 is not among the highest chemical intake rates, it is measured at the highest blood and plasma concentrations for all three ages. This therefore models behaviour of a chemical that chronically accumulates over time and persists in the body. Conversely, cyanuric acid was responsible for extremely high intake levels through both mouthing ingestion and dermal absorption, and continued to appear in the body as the chemical with the greatest concentration in urine for all three ages. This means that although cyanuric acid is consumed and absorbed at the highest rates, it is also consistently excreted at the highest levels, limiting its persistence and potential for toxicity.

Although the parameters of the simulated environment were not shared, we can make some predictions. With cyanuric acid measured at the highest intake levels for both dermal permeation and mouthing ingestion, its presence in this simulated environment is clear.

Cyanuric acid is a common chemical used in swimming pools used to limit the amount of chlorine required[17]; and individuals who swim are mainly exposed through consistent skin contact with the water, and exposure to water through an open mouth (especially for children).[18] Therefore it is reasonable to assume that the reason dermal permeation and mouthing ingestion rates were dominated by cyanuric acid levels is due to environmental exposure through swimming pool water.

Section VI: Future work

After a unique multidisciplinary introduction to python programming, the vast array of feature expansion options is recognized and motivating. Firstly, the introduction of chemical informatic libraries such as RDKit and PubChemPy would allow for feature expansion of additional chemical properties. With the chemical name and CAS number given in the dataset, these could be used to query large databases like PubChem to automatically add properties such as the SMILES, LogP, molecular weight, phase change temperatures and vapour pressure to the dataset. If automated, this could extend the number of variables used to evaluate toxicokinetics of many more molecules, enhancing the model's capabilities and knowledge base. Also related to automation, if extended from an introductory investigatory project to a larger recurring project, automation could allow for robust standardization of procedures. If datasets with consistent formatting are utilized, scripts could be run to perform quantitative analysis on the entire dataset; with if statements used to generate reports based on predetermined values like chemical concentration. While many of the techniques utilized in this notebook could be executed in other software's, this project developed familiarity and confidence with Python, and enlightened me to future feature expansions using automation and machine learning, enhancing my computational toolbox as a future chemist.

Section VII: References

(1) Flowers, P.; Theopold, K.; Langley, R. *1.1: Chemistry- The Central science*. Chemistry LibreTexts. https://chem.libretexts.org/Courses/College_of_the_Canyons/Chem_201%3A_General_Chemistry_I_OER/01%3A_Matter_and_Measurements/1.01%3A_Chemistry-_The_Central_science.
(2) Gressling, T. *Data Science in Chemistry: Artificial Intelligence, Big Data, Chemometrics and Quantum Computing with Jupyter*; De Gruyter graduate; Walter de Gruyter: Berlin Boston, 2021.
(3) Tanemura, K. A.; Sierra-Costa, D.; Merz, K. M. Jr. *Python for Chemists*; ACS In Focus; American Chemical Society, 2021. https://doi.org/10.1021/acsinfocus.7e5030.
(4) Vargas-Hernández, R. A. CHEM 4PB3 – Computational Models for Electronic Structure and Chemical Bonding 2022 Winter Term.
(5) Okeme, J. O. CHEM 4SCS – Sustainable Chemistry: Analysis and Regulation 2023 Winter Term.
(6) Schulte, P. A.; McKernan, L. T.; Heidel, D. S.; Okun, A. H.; Dotson, G. S.; Lentz, T. J.; Geraci, C. L.; Heckel, P. E.; Branche, C. M. Occupational Safety and Health, Green Chemistry, and Sustainability: A Review of Areas of Convergence. *Environ. Health* **2013**, *12*, 31. https://doi.org/10.1186/1476-069X-12-31.
(7) Jia, X.; Wang, T.; Zhu, H. Advancing Computational Toxicology by Interpretable Machine Learning. *Environ. Sci. Technol.* **2023**, *57* (46), 17690–17706. https://doi.org/10.1021/acs.est.3c00653.
(8) Shanmugam, P. S. T.; Sampath, T.; Jagadeeswaran, I.; Bhalerao, V. P.; Thamizharasan, S.; V., K.; Saha, J. Chapter 12 - Toxicokinetics. In *Biocompatibility Protocols for Medical Devices and Materials*; Timiri Shanmugam, P. S., Sampath, T., Jagadeeswaran, I., Eds.; Academic Press, 2023; pp 175–186. https://doi.org/10.1016/B978-0-323-91952-4.00012-X.
(9) Hemmerich, J.; Ecker, G. F. In Silico Toxicology: From Structure–Activity Relationships towards Deep Learning and Adverse Outcome Pathways. *WIREs Comput. Mol. Sci.* **2020**, *10* (4), e1475. https://doi.org/10.1002/wcms.1475.
(10) Li, L.; Sangion, A.; Wania, F.; Armitage, James. M.; Toose, L.; Hughes, L.; Arnot, J. A. Development and Evaluation of a Holistic and Mechanistic Modeling Framework for Chemical Emissions, Fate, Exposure, and Risk. *Environ. Health Perspect.* **2021**, *129* (12). https://doi.org/10.1289/EHP9372.
(11) Anand, S. S.; Mehendale, H. M. Volatile Organic Compounds (VOC). In *Encyclopedia of Toxicology (Second Edition)*; Wexler, P., Ed.; Elsevier: New York, 2005; pp 450–455. https://doi.org/10.1016/B0-12-369400-0/01015-2.
(12) *What Percentage of the Body is Water?*. DexaFit. https://www.dexafit.com/blog2/what-percentage-of-the-body-is-water.
(13) Schrier Stanley L. Water Content of Erythrocytes. *N. Engl. J. Med.* **1974**, *291* (10), 526–527. https://doi.org/10.1056/NEJM197409052911011.
(14) Mathew, J.; Sankar, P.; Varacallo, M. Physiology, Blood Plasma. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2024.
(15) *29.8: Urine Composition and Function*. Chemistry LibreTexts. https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Map%3A_Fundamentals_o

f_General_Organic_and_Biological_Chemistry_(McMurry_et_al.)/29%3A_Body_Fluids/29.08%3A_Urine_Composition_and_Function.

(16) *12.4: Modifiers of Biotransformation*. Chemistry LibreTexts. https://chem.libretexts.org/Bookshelves/Environmental_Chemistry/Toxicology_MSDT/6%3A_Principles_of_Toxicology/Section_12%3A_Biotransformation/12.4%3A_Modifiers_of_Biotransformation (accessed 2024-04-26).

(17) Chen, Z.; Su, Y.; Chen, J.; Li, Z.; Wang, T. Study on the Health Risk of Cyanuric Acid in Swimming Pool Water and Its Prevention and Control Measures. *Front. Public Health* **2024**, *11*. https://doi.org/10.3389/fpubh.2023.1294842.

(18) US EPA, O. *Swimmer Exposure Assessment Model (SWIMODEL)*. https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/swimmer-exposure-assessment-model-swimodel (accessed 2024-04-26).