


2024-06-07

# volcalc: Calculate predicted volatility of chemical compounds

Kristina Riemer 

Communications & Cyber Technologies, University of Arizona

Eric R. Scott 

Communications & Cyber Technologies, University of Arizona

Laura Meredith 

School of Natural Resources and the Environment, University of Arizona

S. Marshall Ledford

School of Natural Resources and the Environment, University of Arizona

## Signatories

### Project team

- Kristina Riemer, author/maintainer, Director of Communications & Cyber Technologies Data Science team at University of Arizona
- Eric Scott, contributor, Scientific Programmer and Educator for Communications & Cyber Technologies Data Science team at University of Arizona

### Contributors

- Assistant Professor Laura Meredith developed the original idea for the `volcalc` R package along with Kristina Riemer and supports continued development by our team.
- PhD student S. Marshall Ledford has been the main user of early versions of `volcalc` and will continue to provide feedback on the package API and documentation.

## Consulted

Tamás Stirling, maintainer of the `webchem` R package (part of rOpenSci), was consulted and confirmed that `volcalc` does not replicate efforts of any similar R packages that we are aware of.

## The Problem

Volatile organic compounds (VOCs) are important components of many biological and chemical processes, yet estimates of compound volatility are not available for most compounds. VOCs readily evaporate under ambient conditions and are important in a number of fields and contexts including plant defense, microbial communication, and indoor air pollution, to name a few. Yet measures of volatility are time consuming to determine experimentally and not available for the vast majority of compounds in chemical information databases.

The `volcalc` R package automates the process of estimating volatility of compounds from their chemical structure. Our package automates these following steps: 1) downloading data on chemical structure, 2) parsing those data to discover chemical functional groups, 3) applying the SIMPOL.1 algorithm (Pankow & Asher, 2008) to predict volatility from functional groups and molecular weight, and assigning compounds a volatility category given a reference environment. This enables fast and easy estimation of volatility for thousands of chemical compounds.

`volcalc` is currently limited to working with compounds present in a single database. We propose expanding its use to include any chemical with structural information—data that is widely available in many chemical information databases. In addition, we will improve testing and documentation to make the package more reliable.

There are currently no other user-friendly, automated tools for predicting compound volatility.

## The proposal

### Overview

Using the support from this proposal, we will take our existing tool for calculating volatility of compounds and expand its usage to make predictions for essentially any compound, and enable researchers in a wide variety of fields to use it with ease. The current version of `volcalc` can estimate volatility only for compounds in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa, 2000). However, by refactoring existing code, we can make the volatility prediction functionality available to any compound with a known structure. Additionally, we plan to improve package infrastructure and documentation (e.g. increased test coverage, continuous integration, vignettes), build the package on r-universe, and submit the package to CRAN to improve discoverability and simplify installation.

### Detail

The current version of `volcalc` was created in 2022 as part of a [data science incubator](#) project in collaboration between Dr. Kristina Riemer and Dr. Laura Meredith at University of Arizona. `volcalc` is developed on GitHub and distributed under an MIT license. `volcalc` is the first project, to our knowledge, to implement the SIMPOL method for predicting chemical vapor pressures and enthalpies of vaporization (Pankow & Asher, 2008) in an R package. `volcalc` has been successfully used to calculate volatility estimates for *all* 19,000+ compounds in the KEGG database with avail-

able masses (L. Meredith et al., *in prep*). To ensure accuracy of the package, the volatility estimates for a subset of the compounds were validated against hand-collected data (Honeker et al., 2021; L. K. Meredith & Tfaily, 2022).

The **first minimum viable product** is a function that can calculate volatility when provided a path to a molfile and a package vignette demonstrating how to couple this with chemical data sources such as the `webchem` package. The main function in `volcalc`, `calc_vol()`, currently works by downloading a molfile from the KEGG API given a KEGG-specific chemical identifier as a starting point. Molfile is an open file format for storing chemical structure data, and various tools exist to translate other standard representations of chemical structure such as SMILES (Weininger, 1988) and InChI (Heller et al., 2013) to molfiles. This translation can be done using `ChemmineOB` which is already a dependency of `volcalc`. SMILES and InChI are both string representations, so adding these as possible inputs will make `volcalc` fit more easily into a `data.frame`-based workflow.

The **second minimum viable product** is a package that has gone through the steps suggested by `usethis::use_release_issue()` and is ready to submit to CRAN.

Refactoring the code in `volcalc` to work with essentially any chemical and preparing the package for wider distribution will make this powerful tool accessible to researchers across a variety of domains.

Project repository: <https://github.com/Meredith-Lab/volcalc>

## Project plan

### Dates

- Project start date: June 1, 2023
- Project end date: May 31, 2024

### Start-up phase

#### Milestone 1: August 1, 2023

- Implement CI with GitHub actions
- Check code coverage with `codecov` package
- Use GitHub Issues or Discussions to brainstorm eventual API (e.g., function names, argument names, how many exported functions)

Estimated work: 20 hours

### Technical delivery

#### Milestone 2: October 1, 2023

- Re-factor `calc_vol()` code to split KEGG download and SIMPOL calculation functionality
- Implement user inputs for environment (soil, water) and volatility categories
- Deprecate arguments and functions appropriately as necessary
- Update documentation to reflect new function usage

Estimated work: 70 hours

### Milestone 3: February 1, 2024

- Create a vignette demonstrating both KEGG usage and more general usage (i.e., providing a path to a molfile) for volatility estimation
- Improve package documentation by adding citations, details, and additional examples where appropriate
- Create a `pkgdown` website for `volcalc`
- Create a `CITATION.cff` file, make a GitHub release, and archive code on Zenodo
- Go through checklist generated by `usethis::use_release_issue()` in preparation for CRAN submission

Estimated work: 70 hours

### Milestone 4: May 1, 2024

- Add functionality to supply other chemical representations besides molfiles as input
- Add to vignette(s) examples of integrating `volcalc` with data sources such as the `webchem` package to estimate volatility for an arbitrary set of compounds (i.e., not from KEGG)

Estimated work: 80 hours

## Other aspects

Dissemination plan:

- After the initial re-factor (milestone 2 above), we plan to share the package with `webchem` contributors via our rOpenSci Slack channel for feedback & suggestions. We will encourage them to share the project with their networks of collaborators as well.
- Near the project conclusion we will:
  - prepare a blog post for <https://datascience.cct.arizona.edu/>
  - prepare a Twitter announcement to share from [@cct\\_datascience](#)
  - prepare a short demonstration video to be published to our [YouTube channel](#) that can be shared on the package README and through social media
  - work with our collaborator Laura Meredith to identify potential domain-specific venues to promote the use of `volcalc` including virtual training workshops, international scientific conferences, email lists, lab website, and social media accounts

Estimated work: 20 hours

## Requirements

### People

Kristina Riemer and Eric Scott will do the coding work for this project. Eric and Kristina both have experience creating R packages, writing tests, and collaborating using GitHub. Eric has experience submitting packages to CRAN and with automation using GitHub Actions.

Eric and collaborators Laura Meredith and S. Marshall Ledford have domain experience working with VOC data and chemoinformatics more generally. S. Marshall Ledford is currently the main user of `volcalc` besides the developers, and will give feedback on future versions.

## Funding

We request funding for the salaries of personnel working on this project.

- Kristina Riemer: 100 hours, \$4,791
- Eric Scott: 160 hours, \$7,475

For a total of \$12,265

Funding timeline:

- After milestone 2 (October 1, 2023), \$6,133
- After technical delivery (May 31, 2024), \$6,132

## Summary

These costs represent a one-time investment to pay for personnel to work on the project.

## Success

### Definition of done

We would consider this project successful when a new version of `volcalc` with the ability to calculate volatility given an arbitrary molfile has been released on GitHub, the code has been archived on Zenodo, and the package has been successfully submitted to CRAN.

### Measuring success

The following can be used to measure success:

- Can the package be installed from GitHub? (yes/no)
- Can the package be installed from r-universe? (yes/no)
- Is there sufficient test coverage? (at least 90%)
- Are there correctness tests for volatility predictions? (yes/no)
- Is the package passing R CMD `check` on Linux, macOS, and Windows using CI? (yes/no)
- Has the package been tested by users other than the developers? (yes/no)
- Does the package have a vignette that is easy to follow? (yes/no)
- Is code archived on Zenodo and a DOI associated with the package citation? (yes/no)
- Can chemical representation types be used as input? (molfile for success; InChI, SMILES, and possibly more for reach goal)

### Future work

- The contribution of some functional groups to volatility were not able to be measured using existing methods. As these analytical methods improve, the algorithm in `volcalc` can be updated.
- `volcalc` only calculates relative volatility under standard conditions; functionality to specify custom environmental conditions could be added.

## Key risks

- The SIMPOL algorithm might not be applicable to *all* compounds, since it was designed to work with volatile organic compounds. In that case, we may need to add a warning to the user that values returned by `volcalc` may not make sense for certain compounds based on molecular weight, for example. We will explore edge cases with our collaborators and develop a plan to guide users to correct and accurate interpretation of data returned by `volcalc` functions.
- `volcalc` has `OpenBabel` as a system dependency (indirectly), which could potentially lead to delays in getting the package to build using GitHub Actions. In the event that this is an issue, we will reach out to the rOpenSci community on Slack for help with this.
- We plan to improve and submit this package to CRAN to increase the possible userbase and its ease of use. This process can be challenging due to technical hurdles, and it is possible (although unlikely) that this package will not ultimately be accepted. If this is the case, the package will still be available for easy installation from [cct-datascience.r-universe.dev](https://cct-datascience.r-universe.dev).

## References

- Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., & Pletnev, I. (2013). InChI - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1). <https://doi.org/10.1186/1758-2946-5-7>
- Honeker, L. K., Graves, K. R., Tfaily, M. M., Krechmer, J. E., & Meredith, L. K. (2021). The volatilome: A vital piece of the complete soil metabolome. *Frontiers in Environmental Science*, 9. <https://doi.org/10.3389/fenvs.2021.649905>
- Kanehisa, M. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Meredith, L. K., & Tfaily, M. M. (2022). Capturing the microbial volatilome: An oft overlooked 'ome'. *Trends in Microbiology*, 30(7), 622–631. <https://doi.org/10.1016/j.tim.2021.12.004>
- Meredith, L., Riemer, K., Geffre, P., Honeker, L., Krechmer, J., Graves, K., Tfaily, M., & Ledford, S. (*in prep*). Automating methods for estimating metabolite volatility.
- Pankow, J. F., & Asher, W. E. (2008). SIMPOL.1: A simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds. *Atmos. Chem. Phys.* <https://doi.org/10.5194/acp-8-2773-2008>
- Weininger, D. (1988). SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1), 31–36.