2025-10-18

ChemistryShinyTools: Streamlining R Shiny Applications for Cheminformatics Research

Saw Simeon
Department of Chemistry, Kasetsart University

ChemistryShinyTools Proposal

This document outlines the proposal for `ChemistryShinyTools`, an open-source R package designed to simplify the development of R Shiny applications for cheminformatics research. The project aims to empower chemists and data scientists to create interactive tools for molecular visualization, QSAR modeling, and drug discovery, promoting R's adoption in the chemistry community.

Executive Summary

Title: ChemistryShinyTools: Streamlining R Shiny Applications for Cheminformatics Research

Proposer: Saw Simeon, PhD, saw.s@ku.th

`ChemistryShinyTools` is an open-source R package designed to simplify the creation of R Shiny applications for chemistry and cheminformatics research, enabling chemists and data scientists to build interactive tools for molecular visualization, QSAR modeling, and drug discovery. The package provides modular Shiny components, machine learning integration, and reproducible workflow templates. Developed by Saw Simeon, a PhD with expertise in R Shiny development and cheminformatics (e.g., publications in Nature Protocols 2023 and Journal of Cheminformatics 2020), this project leverages a proven track record of delivering user-focused tools to advance the R ecosystem.

Goals

1. **Promote R in Chemistry**: Establish R as a go-to tool for cheminformatics by providing accessible Shiny modules for molecular data analysis.

2. **Support Infrastructure**: Develop reusable, modular tools to reduce development time for chemistry-focused Shiny applications.

3. **Increase User Adoption**: Lower barriers for chemists through comprehensive documentation, tutorials, and pre-built components.

4. **Foster Collaboration**: Engage the R and chemistry communities via GitHub, encouraging contributions and feedback.

Methods

The package will be developed using R, leveraging packages like `shiny`, `ggplot2`, `caret`, and `dplyr`, with optional integration of `rcdk` and `fingerprint` for molecular data processing. Key tasks include:
- **Development**: Create modular Shiny components (e.g., for visualizing SMILES strings, molecular descriptors) and machine learning wrappers (e.g., for QSAR modeling)
- **Documentation**: Produce vignettes and tutorials using Quarto.
- **Collaboration**: Host the package on GitHub (github.com/sawsimeon/ChemistryShinyTools), using Git for version control and soliciting community input via issues and pull requests.
- **Testing**: Implement unit tests with testthat and ensure CRAN compliance using devtools.
- **Outreach**: Share the package on X and R forums (e.g., RStudio Community) to engage users.

Expected Outcomes

- **Increased Adoption**: Chemists and cheminformaticians adopt `ChemistryShinyTools` for rapid development of interactive tools, to enhance R's presence in chemistry research.

- **Improved Productivity**: Reusable modules reduce Shiny app development time.
- **Community Engagement**: At least 10 active contributors on GitHub within 12 months
- **Reproducible Research**: Standardized templates and documentation promote reproducible workflows.

Deliverables

1. **R Package**: `ChemistryShinyTools` on CRAN and GitHub, including:

- Shiny modules for molecular visualization (e.g., PCA plots).

- Machine learning wrappers for QSAR and virtual screening.

- Chemical data processing utilities (e.g., SMILES parsing).

2. **Documentation**: A vignette and tutorials, rendered to HTML/PDF via GitHub Pages.

3. **Example App**: A sample Shiny app demonstrating molecular data visualization.

4. **Community Resources**: A `CONTRIBUTING.md` file and X posts to engage the R and chemistry communities.

Budget

Estimated needs include:
- **Development Time**: 6 months part-time (20 hours/week) for coding, testing, and documentation.
- **Software/Tools**: Free/open-source tools (R, Quarto, GitHub, rcdk); optional cloud hosting
- **Total Estimate**: ~$1,000–$2,000 Funding will support development, documentation, and community engagement, ensuring alignment with R Consortium goals.

Timeline

- **Month 1–2**: Develop core Shiny modules and machine learning wrappers; set up GitHub repo.
- **Month 3–4**: Create documentation, vignettes, and example app; begin community outreach on X.
- **Month 5**: Test package, ensure CRAN compliance, and solicit feedback via GitHub.
- **Month 6**: Submit to CRAN and ISC; present at an Uppsala Open Science group.

Conclusion

`ChemistryShinyTools` will empower chemists to leverage R's statistical computing power through intuitive Shiny applications, advancing the R Consortium's mission to promote R, support infrastructure, and foster collaboration. The proposer's expertise in R Shiny, cheminformatics, and reproducible research ensures a high-impact project that bridges academia and the R community.

Signatories

This section outlines the individuals involved in the `ChemistryShinyTools` project, detailing the project team responsible for delivery, contributors who have supported the proposal, and those consulted for feedback. The broad support from the R and cheminformatics communities underscores the project's alignment with community needs and its potential for successful implementation.

Project Team

The core team responsible for delivering `ChemistryShinyTools` consists of individuals with complementary expertise in R Shiny development, cheminformatics, and open-source software development:

- **Saw Simeon, PhD (Principal Developer)**: Proposer and lead developer with extensive experience in R Shiny application development and cheminformatics. Published in *Nature Protocols* (2023) and *Journal of Cheminformatics* (2020), Saw specializes in creating user-focused tools for molecular data analysis and visualization. Responsibilities include package development, documentation, and community outreach.

Contributors

- **Open Science Group at Uppsala University**: Provided early feedback on the proposal's open-source collaboration model and suggested strategies for community engagement via GitHub.

Consulted

The following individuals and groups were consulted to ensure the proposal addresses community needs and aligns with the R Consortium's mission:

- **R Consortium Infrastructure Steering Committee (ISC)**: Preliminary discussions were held with ISC member to ensure the proposal supports R's adoption in scientific research.
- **GitHub Community**: Early ideas were shared on GitHub discussions to gauge interest and solicit suggestions from R and cheminformatics developers.

(**?**; Wickham 2016)

The Problem

Cheminformatics research relies heavily on the ability to process, analyze, and visualize complex molecular data, yet creating user-friendly, interactive tools for these tasks remains a significant challenge. Developing R Shiny applications tailored for chemistry and cheminformatics is time-intensive, requiring expertise in both R programming and domain-specific knowledge of molecular data handling. This complexity creates barriers for chemists and data scientists who lack advanced programming skills, limiting their ability to leverage R's powerful statistical and visualization capabilities for tasks like molecular visualization, quantitative structure-activity relationship (QSAR) modeling, and drug discovery.

What the Problem Is

The primary issue is the lack of accessible, modular, and reusable tools for building R Shiny applications specifically designed for cheminformatics. Existing workflows often require developers to create custom Shiny components from scratch, integrating specialized packages like `rcdk` or `fingerprint` for molecular data processing. This process is not only time-consuming but also prone to errors and inconsistencies, as there are no standardized templates or pre-built modules tailored to cheminformatics needs.

Who It Affects

This problem affects a wide range of stakeholders in the chemistry and data science communities: - **Chemists and Cheminformaticians**: Researchers who need intuitive tools to visualize and analyze molecular data but lack the programming expertise to build custom Shiny applications. - **Data Scientists**: R users who work with chemical data but spend excessive time developing bespoke solutions due to the absence of modular, chemistry-focused Shiny components. - **Academic and Industry Researchers**: Teams conducting drug discovery or QSAR modeling who require reproducible, interactive tools to streamline their workflows. - **R Community**: Developers seeking to expand R's adoption in cheminformatics but facing high entry barriers due to the complexity of integrating domain-specific tools.

Why It Is a Problem

The absence of standardized, reusable Shiny components for cheminformatics hinders research efficiency and reproducibility. Chemists without advanced R programming skills are forced to rely on less flexible, proprietary software or manual workflows, which are often costly or lack the statistical rigor of R. For data scientists, the need to repeatedly build custom Shiny applications for molecular visualization or machine learning tasks (e.g., QSAR modeling) results in duplicated

effort and inconsistent outputs. This inefficiency slows down research progress, limits collaboration, and restricts R's adoption in the chemistry domain, despite its potential to be a leading tool for cheminformatics.

What Solving the Problem Will Enable

Addressing this problem will: - **Empower Chemists**: Provide accessible tools for non-programmers to create interactive Shiny applications for molecular data analysis, visualization, and drug discovery. - **Enhance Productivity**: Reduce development time through reusable, modular Shiny components, allowing researchers to focus on scientific discovery rather than coding. - **Promote Reproducibility**: Standardize workflows with documented, open-source tools, ensuring consistent and shareable results across research teams. - **Expand R's Reach**: Increase R's adoption in cheminformatics by making it easier to build chemistry-focused applications, aligning with the R Consortium's mission to promote R. - **Foster Collaboration**: Encourage community contributions through an open-source package, creating a sustainable ecosystem for cheminformatics tools in R.

Summary of Existing Work and Previous Attempts

Several R packages support cheminformatics tasks, such as `rcdk` for molecular descriptor calculations, `fingerprint` for molecular similarity analysis, and `ChemmineR` for chemical data processing (**?**; **?**; **?**). However, these packages focus on data processing rather than providing modular components for building interactive Shiny applications. While general-purpose Shiny packages like `shiny` and `shinydashboard` enable web application development, they lack chemistry-specific features, such as SMILES string visualization or QSAR model integration. Existing cheminformatics Shiny applications, such as those described in *Nature Protocols* (2023), are often project-specific and not reusable, requiring developers to start from scratch for new applications (**?**). No comprehensive, open-source R package currently exists to streamline the creation of chemistry-focused Shiny applications with modular components and extensive documentation.

Changes to R Itself

This proposal does not involve changes to R itself, so no letter of support from an R Core member is required.

(Wickham 2016; **?**; **?**; **?**; **?**)

The Proposal

This proposal outlines the development of `ChemistryShinyTools`, an open-source R package designed to streamline the creation of R Shiny applications for cheminformatics research. By providing modular Shiny components, machine learning integration, and comprehensive documentation, the project addresses the challenge of building user-friendly, chemistry-focused interactive tools. Below, we detail the approach, concrete actions, timeline, and potential failure modes, emphasizing benefits to the R and cheminformatics communities.

Overview

`ChemistryShinyTools` aims to simplify the development of R Shiny applications for cheminformatics by offering reusable modules for molecular visualization, QSAR modeling, and chemical data processing. The package will enable chemists and data scientists to create interactive tools without extensive programming expertise, promoting R as a go-to platform for cheminformatics research. Key benefits

to the R community include: - **Increased R Adoption**: Lowering barriers for chemists to use R for interactive data analysis. - **Enhanced Productivity**: Reusable modules reduce development time for Shiny applications. - **Community Collaboration**: An open-source model fosters contributions via GitHub, aligning with the R Consortium's mission. - **Reproducible Research**: Standardized templates and documentation ensure consistent, shareable workflows.

Concrete Actions and Timeline

- **Month 1–2**: Develop core Shiny modules (e.g., SMILES visualization, PCA plots) and machine learning wrappers for QSAR modeling. Set up GitHub repository (github.com/sawsimeon/ChemistryShinyTools) with `CONTRIBUTING.md`. Estimated effort: 160 hours.
- **Month 3–4**: Create documentation using Quarto, including vignettes and tutorials. Build a sample Shiny app demonstrating molecular visualization. Begin outreach on X and RStudio Community. Estimated effort: 160 hours.
- **Month 5**: Implement unit tests with `testthat`, ensure CRAN compliance using `devtools`, and solicit community feedback via GitHub issues. Estimated effort: 80 hours.
- **Month 6**: Finalize package, submit to CRAN, and present at Uppsala Open Science group. Estimated effort: 80 hours.

Total development time: 6 months part-time (20 hours/week).

Potential Failure Modes and Recovery

- **Failure Mode**: Limited community adoption due to insufficient outreach.
  - **Recovery**: Intensify promotion on X, RStudio Community, and chemistry forums; collaborate with R Consortium for visibility.
- **Failure Mode**: Technical challenges in integrating `rcdk` or `fingerprint` for molecular processing.
  - **Recovery**: Use alternative packages (e.g., `ChemmineR`) or simplify functionality to ensure core deliverables are met.
- **Failure Mode**: Delays in CRAN submission due to compliance issues.
  - **Recovery**: Leverage `devtools` for early compliance checks and seek feedback from R community experts.

Detail

Minimum Viable Product

The minimum viable product (MVP) for `ChemistryShinyTools` will include: - A Shiny module for visualizing SMILES strings as 2D molecular structures. - A machine learning wrapper for QSAR modeling using `caret`. - Basic documentation with one vignette and a tutorial hosted on GitHub Pages. - A sample Shiny app demonstrating molecular visualization. This MVP delivers immediate value by enabling chemists to create simple interactive tools for molecular analysis, addressing the core problem of time-intensive Shiny development.

Architecture

The package will follow a modular architecture: - **Shiny Modules**: Reusable components for molecular visualization (e.g., PCA plots, SMILES rendering) and data input/output. - **Machine Learning Layer**: Wrappers for QSAR modeling and virtual screening, built on `caret` and integrated with `rcdk` for molecular descriptors. - **Data Processing Utilities**: Functions for parsing SMILES strings and computing molecular properties, leveraging `rcdk` and `fingerprint`. - **Documentation Framework**: Quarto-based vignettes and tutorials, rendered to HTML/PDF via GitHub Pages. - **Testing and Deployment**: Unit tests with `testthat` and CRAN-compliant structure using `devtools`. The package will be hosted on GitHub, with version control via Git and community contributions managed through issues and pull requests.

Assumptions

- Chemists and data scientists need accessible tools to build Shiny applications for cheminformatics.
- Existing R packages (`rcdk`, `fingerprint`, `caret`) are sufficient for core cheminformatics tasks.
- The R and chemistry communities will engage with an open-source package via GitHub and X.
- A 6-month part-time development timeline is feasible with the proposed budget ($1,000–$2,000). If these assumptions are invalidated (e.g., low community interest), the project will pivot to focus on niche use cases or alternative outreach channels (e.g., academic conferences).

External Dependencies

- **R Packages**: `shiny`, `ggplot2`, `caret`, `dplyr` (core functionality); `rcdk`, `fingerprint`, `ChemmineR` (molecular processing); `testthat`, `devtools` (testing and CRAN compliance); `quarto` (documentation).
- **Services**: GitHub for hosting and collaboration; GitHub Pages for documentation hosting; optional cloud hosting for the sample app.
- **Community Platforms**: X and RStudio Community for outreach and feedback. These dependencies are stable, open-source, and widely used, minimizing risks to project delivery.

(**?**; Wickham 2016; **?**; **?**; **?**)

Project Plan

This document outlines the project plan for `ChemistryShinyTools`, an open-source R package to streamline the development of R Shiny applications for cheminformatics research. The plan covers the start-up phase, technical delivery, dissemination strategies, and budget allocation, ensuring alignment with the R Consortium's goals of promoting R, supporting infrastructure, and fostering collaboration. The timeline spans 6 months, with clear milestones and deliverables to ensure timely completion.

Start-up Phase

The start-up phase focuses on establishing the foundation for `ChemistryShinyTools` to ensure smooth development and community engagement.

- **Setting Up Collaboration Platform**:
  - Create a GitHub repository (`github.com/sawsimeon/ChemistryShinyTools`) for version control and community contributions.

  - Develop a `CONTRIBUTING.md` file outlining guidelines for contributors, including coding standards and pull request processes.
  - Set up GitHub Issues for feature requests and bug tracking, and a discussion board for community feedback.
  - **Timeline**: Month 1 (Weeks 1–2).
- **License Decisions**:
  - Adopt the MIT License to ensure the package is open-source, accessible, and compatible with CRAN submission requirements.
  - Document the license in the repository's `LICENSE` file and package `DESCRIPTION`.
  - **Timeline**: Month 1 (Week 1).
- **Reporting Framework**:
  - Establish a quarterly reporting schedule to update the R Consortium on progress via email and blog posts.
  - Create a project wiki on GitHub to document milestones, progress, and community engagement efforts.
  - Schedule presentations at R Consortium ISC meetings and the Uppsala Open Science group to share updates.
  - **Timeline**: Month 1 (Weeks 1–4).

Technical Delivery

The technical delivery phase involves developing, testing, and deploying `ChemistryShinyTools` with clear milestones and target dates to ensure accountability.

- **Month 1–2: Core Development (Target: December 18, 2025)**
  - Develop core Shiny modules for molecular visualization (e.g., SMILES string rendering, PCA plots) using `shiny` and `ggplot2`.
  - Create machine learning wrappers for QSAR modeling using `caret`, integrated with `rcdk` for molecular descriptors.
  - Implement data processing utilities for SMILES parsing and molecular property calculations using `rcdk` and `fingerprint`.
  - Set up package structure with `devtools` for CRAN compliance.
  - **Deliverable**: Initial codebase with functional Shiny modules and QSAR wrappers in the GitHub repository.
- **Month 3–4: Documentation and Example App (Target: February 18, 2026)**
  - Write comprehensive documentation using Quarto, including a vignette and tutorials for molecular visualization and QSAR modeling.
  - Develop a sample Shiny app demonstrating molecular data visualization and basic QSAR functionality.
  - Begin unit testing with `testthat` to ensure module reliability.
  - **Deliverable**: Vignette, tutorials, and sample app hosted on GitHub Pages; initial test suite completed.
- **Month 5: Testing and Community Feedback (Target: March 18, 2026)**

  - Finalize unit tests and ensure CRAN compliance using `devtools` and `R CMD check`.
  - Solicit community feedback via GitHub Issues and X posts, incorporating suggestions into the package.
  - Refine documentation based on user feedback.
  - **Deliverable**: CRAN-ready package with tested modules and updated documentation.

- **Month 6: Deployment and Outreach (Target: April 18, 2026)**
  - Submit `ChemistryShinyTools` to CRAN and the R Consortium ISC for review.
  - Publish a final vignette and tutorials on GitHub Pages.
  - Present the project at the Uppsala Open Science group and share a completion blog post on the R Consortium blog.
  - **Deliverable**: Published package on CRAN, final documentation, and presentation slides.

Other Aspects

To maximize accessibility and impact, `ChemistryShinyTools` will be developed and shared with the following strategies:

- **Open-Source License**: The package will use the MIT License, ensuring broad accessibility and compatibility with CRAN and community contributions.
- **Code Hosting**: The GitHub repository (`github.com/sawsimeon/ChemistryShinyTools`) will serve as the central hub for code, documentation, and collaboration, with Git for version control and Issues for feedback.
- **Publicity**:
  - **Announcement Post**: Publish an announcement on the R Consortium blog and X in Month 2 to introduce the project and invite contributors.
  - **Delivery Blog Post**: Share a completion post on the R Consortium blog in Month 6, detailing outcomes and usage instructions.
  - **Social Media**: Regular updates on X ((**?**)) and RStudio Community forums to engage chemists and R users (Months 2–6).
  - **UseR!**: Submit an abstract to UseR! 2026 to present `ChemistryShinyTools` and demonstrate its applications (Month 5).
  - **ISC Meetings**: Provide quarterly updates at R Consortium ISC meetings to ensure alignment with community goals (Months 3 and 6).
- **Quarterly Content**: Share progress updates on the R Consortium blog every 3 months, covering development milestones, community feedback, and use cases.

Budget & Funding Plan

The estimated budget for `ChemistryShinyTools` is $1,000–$2,000, primarily allocated to labor costs for development, testing, and documentation. The budget is tied to the following milestones:

- **Milestone 1: Core Development (Months 1–2, $800)**
  - **Work**: Develop Shiny modules, QSAR wrappers, and data processing utilities; set up GitHub repository and license.
  - **Outcome**: Functional codebase with initial modules and repository infrastructure.

  - **Funding**: Covers 160 hours of part-time development (20 hours/week).

- **Milestone 2: Documentation and Example App (Months 3–4, $800)**
  - **Work**: Create Quarto-based documentation, tutorials, and a sample Shiny app; begin testing and outreach.
  - **Outcome**: Vignette, tutorials, and app hosted on GitHub Pages; initial test suite.
  - **Funding**: Covers 160 hours of part-time work for documentation and app development.

- **Milestone 3: Testing, Deployment, and Outreach (Months 5–6, $400)**
  - **Work**: Finalize testing, ensure CRAN compliance, incorporate community feedback, submit to CRAN, and present at Uppsala Open Science group.
  - **Outcome**: CRAN-published package, finalized documentation, and community engagement.
  - **Funding**: Covers 160 hours of part-time work for testing, refinement, and outreach.

- **Additional Costs**:
  - **Software/Tools**: Free/open-source tools (R, Quarto, GitHub, `rcdk`, `fingerprint`) will be used, minimizing costs.
  - **Cloud Hosting**: Optional cloud hosting for the sample app (e.g., ShinyApps.io, ~$100–$200, if needed).
  - **Total Estimate**: $1,000–$2,000, excluding indirect costs (e.g., travel, hardware), which are not covered per ISC guidelines.

The budget ensures efficient use of funds, focusing on labor to deliver a high-quality, community-driven package. The ISC may vet fund usage, and any questions will be addressed promptly.

(**?**; Wickham 2016; **?**; **?**; **?**)

The timeline section is under construction and will be updated soon.

Success

Definition of done

Measuring success

Future work

The success criteria section is under construction and will be updated soon.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.