CRYSS: a Data Analysis System for High-Throughput Crystallization Screening

cryss.co.uk, 03-SEP-2026

CRYSS is an open-source insight generator for partitioned equilibrium datasets, designed to convert complex multicomponent solubility behaviour into mechanistic, decision-ready purification models.

Executive Summary

Introduction

Chemical synthesis almost always produces side products, making purification an essential part of process development. In multi-step pharmaceutical synthesis, even high step yields accumulate into substantial overall loss: a ten-step sequence at 90% per step delivers only about 35% overall yield. In practice, synthesis is an exercise in impurity removal.

Crystallization is one of the most attractive purification methods when the target molecule is a solid. It is operationally simple, scalable, and cost-effective. Yet optimizing a crystallization process is non-trivial: one must maximize space-time yield while meeting stringent quality targets. This requires controlling polymorphs, crystal growth kinetics, terminal solid-solution behaviour, and supersaturation management via strategies such as seeding, cooling, evaporation, or antisolvent addition.

Before any kinetic optimization begins, however, the primary screening question is simple: which systems provide the highest selective removal of impurities and therefore the highest achievable yield of the target molecule? Answering this requires exploration of solvents, solvent mixtures, and, where chemistry permits, derivatization or salt formation. These choices introduce categorical variables and rapidly increase the combinatorial burden of screening.

Parallel small-reactor blocks with stirring and thermal cycling provide an efficient way to screen purification behaviour across many systems. The challenge becomes how to interpret the results. CRYSS is designed specifically to transform equilibrium-partitioned data into a mechanistic ranking of candidate systems.

CRYSS Workflow Overview

At its simplest, the method consists of the following steps (see Figure 1):

1) Equilibrate the candidate mixture (which may include different salt forms) with various solvents or solvent mixtures in a closed system to avoid solvent loss. Achieving equilibrium is essential, so thermal cycling, agitation, and sufficient time are required.
2) Determine the mass fraction of the mixture in solution and solid phases at equilibrium. If performed carefully, only calibrated analytics of the solution phase are needed: knowing the initial mass and the dissolved mass immediately gives the solid mass by difference. Avoiding solvent loss is critical, as evaporation artificially inflates the apparent dissolved

mass. Alternatively, separating solid and solution and applying area % analysis yields the same information via the lever rule, though this is more difficult at small scale.

3) Record each experiment in an Excel file, including:
    a. starting mixture composition (x_c),
    b. solid phase composition (x_s),
    c. solution phase composition (x_l),
    d. experiment or mixture ID.
4) Feed the dataset into CRYSS, which ranks each experiment by the predicted maximum achievable yield (R_{max}) once optimal process conditions are applied. CRYSS identifies the systems worth further development.

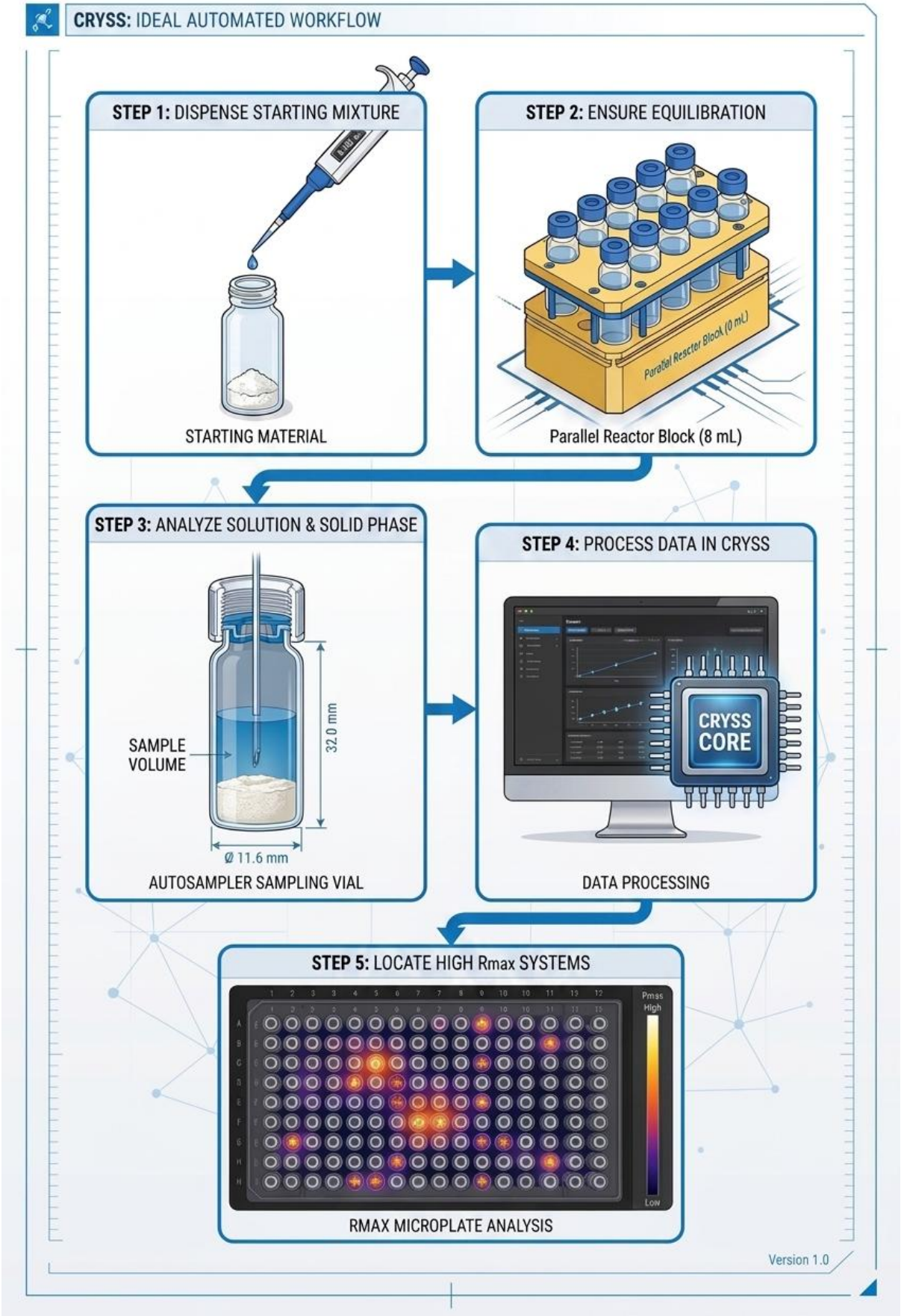


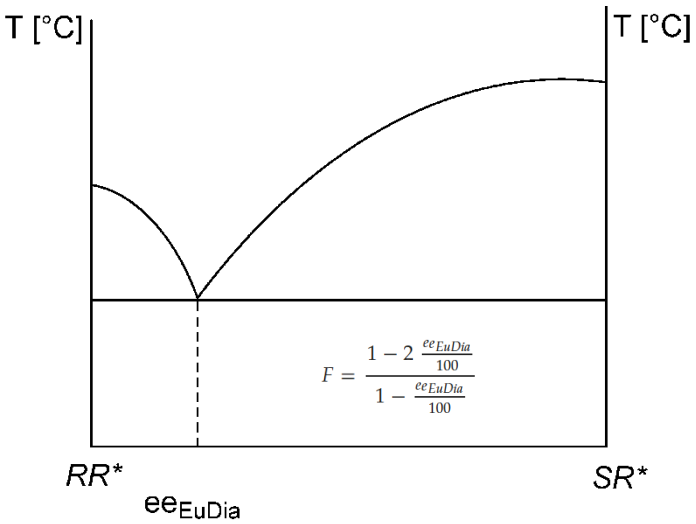
Figure 1: CRYSS enables a highly automated and consistent workflow

Core Principle

1) Binary System

In a binary system (e.g., diastereomeric salt resolution), if both components are present in the solid at equilibrium, the solution phase must have the eutectic composition (x_{eu}). Knowing x_{eu} allows straightforward calculation of the maximum possible recovery of pure material:

$$R_{max} = \frac{(0.5 - x_{eu})}{(1 - x_{eu})}$$



This relationship is familiar from purification of enantioenriched systems.[1] However, multicomponent systems behave differently: the solution phase only has the true eutectic composition at R_{max}, not throughout the entire recovery range.

2) Multicomponent Systems

In multicomponent mixtures, solubility behaviour is encoded in how each component's level changes as a function of equilibrium partitioning at different recovery points R (see Figure 2). For example, if 1 g of mixture is equilibrated with a given solvent volume and 0.1 g dissolves, then 0.9 g remains in the solid phase: i.e. R = 90% recovery. R_{max} is the recovery corresponding to the maximum physical yield of pure material, and achieving it requires an optimal solvent volume. Using more solvent than this reduces yield.

[1] Based on the melting binary phase diagram of diastereomeric mixtures, Ács et al. calculated that, based on the knowledge of the eutectic composition of the binary melting phase diagram, it is possible to calculate the resolvability (F) of the crystalline precipitating diastereomer of the given racemic compound with the given resolving agent. This is a generally applicable method if, after optimization, diastereomeric mixtures have been obtained, and their melting binary phase diagram can be plotted; Pálovics, E.; Madarász, J.; Pokol, G.; Fogassy, E.; Bánhegyi, D.F. Economic Separations of Organic Acidic or Basic Enantiomeric Mixtures—A Protocol Suggestion. *Int. J. Mol. Sci.* **2023**, *24*, 846. https://doi.org/10.3390/ijms24010846

Crucially, all information needed to identify a good purification system is encoded in the analytical profiles across the range R = 100% → R_{max}. This is analogous to diastereomer screening, but requires more sophisticated analysis. CRYSS provides exactly that.

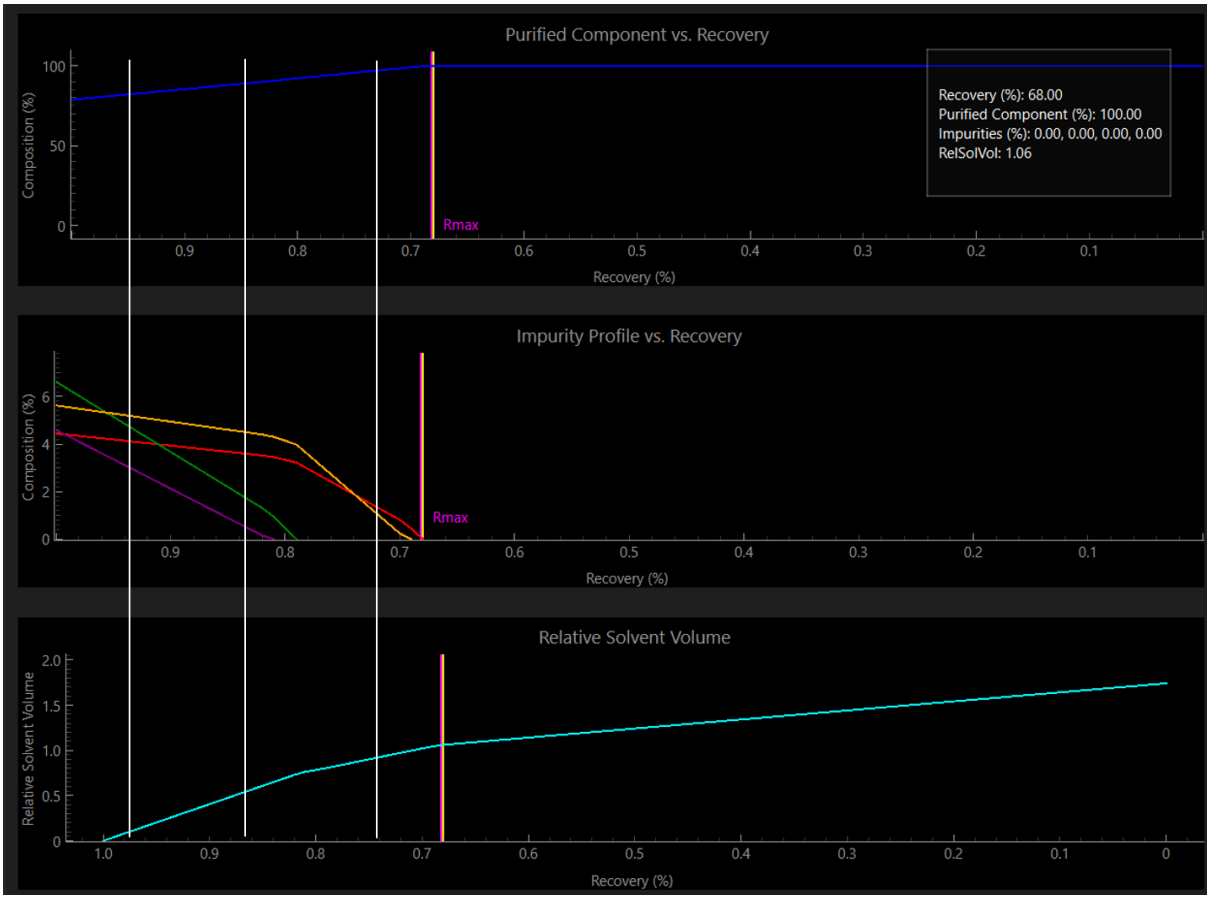


Figure 2 NODES simulation of a 5-component system; the 3 white lines these slices at different Recoveries (R) which correlate with having a different solvent volume (bottom plot) and result in variation of the purity of the target compound (top plot); note the way the impurities vary in a complex way due to their underlying PhysChem parameters. The maximum yield for the system when optimized is 68% physical recovery of pure component 1 (R_{max} = 68%), this the only point in the system where the liquor compositions is the eutectic composition, as distinct from a binary system where the liquors have the eutectic composition throughout the range R = 100% - R_{max}.

3) Understanding Multicomponent Eutectics

In an n-component system:

- The most soluble portion is a eutectic containing all n components.
- Once dissolved, the next most soluble portion is a eutectic of (n–1) components.
- This continues until the least soluble portion remains.

Typically, the least soluble component is the desired product—but occasionally a recalcitrant impurity is least soluble, leading to purification of the wrong species. This is disastrous in practice.

Because the starting composition is known, applying even a single solvent aliquot yields a composition slice at a specific recovery point (white lines in Figure 2). Using a

mechanistic equilibrium model, CRYSS fits the data and projects forward to calculate R_{max}.

The power of the method is its scalability: CRYSS can process large datasets automatically, ranking systems from high-yielding to poor. This enables standardized, rapid triage of large screening arrays and selection of the most promising systems for optimization.

In principle, if the fitted parameters are accurate, a single well-constructed experiment is sufficient to estimate R_{max}. Additional solvent volumes and replicates strengthen the dataset and improve confidence.

Performance of CRYSS

Using the NODES App, we can generate arbitrarily large in silico datasets that act as proxies for real experimental data obtained from the protocol described earlier. This allows us to choose sensible slice points within the critical composition space between R = 100% → R_{max}. Each slice point corresponds to a specific solvent volume in a real experiment. A single slice typically lies near the middle of this range; multiple slices are distributed evenly, with slight randomization to minimize bias.

NODES can generate mathematically closed, precise equilibrium data using different starting compositions, different PhysChem parameter sets, and optional 1:1 addition compounds. These datasets are internally self-consistent, and—critically—we know the true R_{max} for each mixture. This makes them ideal for testing CRYSS.

Beyond perfect numerical data, NODES also includes options to introduce realistic experimental error:

1) Analytical error

    - Capped at 2%
    - Applied with a normal distribution across all components in the initial mixture, solid phase, and liquor phase.

2) Systematic sampling error

    - Simulates cross-contamination of solid into the liquor measurement and liquor into the solid measurement.
    - Models the most common real-world source of error during sampling.

We generated a dataset of 50 mixtures, each with 1–3 slices (proxies for solvent volumes). The files are:

- 260902_1814_Vol01.xlsx
- 260902_1814_Vol01_err5pc.xlsx
- 260902_1814_Vol02.xlsx

- 260902_1814_Vol02_err5pc.xlsx
- 260902_1814_Vol03.xlsx
- 260902_1814_Vol03_err5pc.xlsx

These datasets are available for review here.

CRYSS Performance Results

Single-Volume Data (1 R-slice)

Testing CRYSS with 260902_1814_Vol01.xlsx yields the performance shown in Figure 3, with:

$R^2 = 0.756$

This level of predictivity is already highly acceptable. Even at a single solvent volume, the scatter plot clearly allows rapid visual identification of the best candidates for further experimentation.

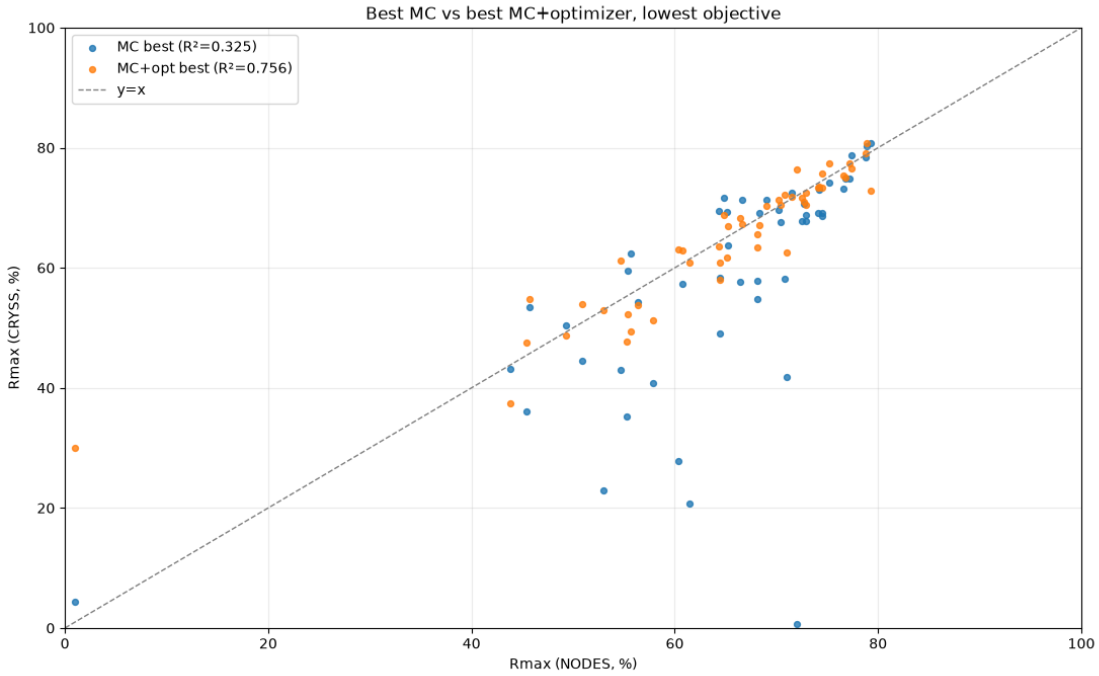


Figure 3: CRYSS run using NODES data from a single solvent volume (one R-slice). The plot compares known NODES R_{max} values with CRYSS-predicted R_{max} values. Blue dots: Monte Carlo simulation only, Orange dots: Hybrid approach (Monte Carlo results fed into an optimizer; best result plotted)

Three-Volume Data (3 R-slices)

Jumping to 260902_1814_Vol03.xlsx, CRYSS achieves:

$R^2 = 0.993$

This is shown in Figure 4, demonstrating near-perfect correlation between true and predicted R_{max} values.

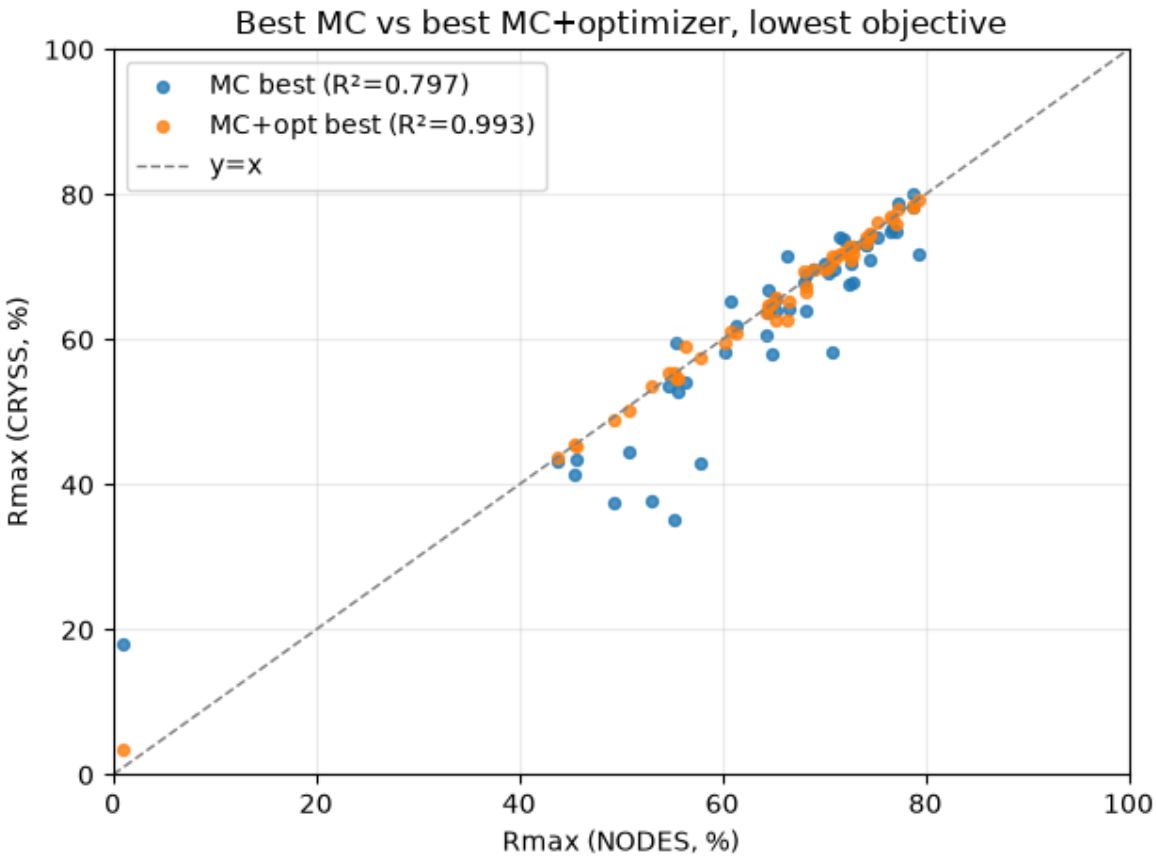


Figure 4: CRYSS run using NODES data from three solvent volumes (three R-slices). Blue dots: Monte Carlo only. Orange dots: Hybrid Monte Carlo + optimizer best result.

Summary Across All Datasets

We repeated this analysis across all NODES test files. The results are summarized in Figure 5. A key observation emerges:

- Experimental error (5%) only becomes visible in the most constrained three-volume datasets.
- For one or two solvent volumes, the inherent noise from fitting dominates, effectively masking the added experimental error.

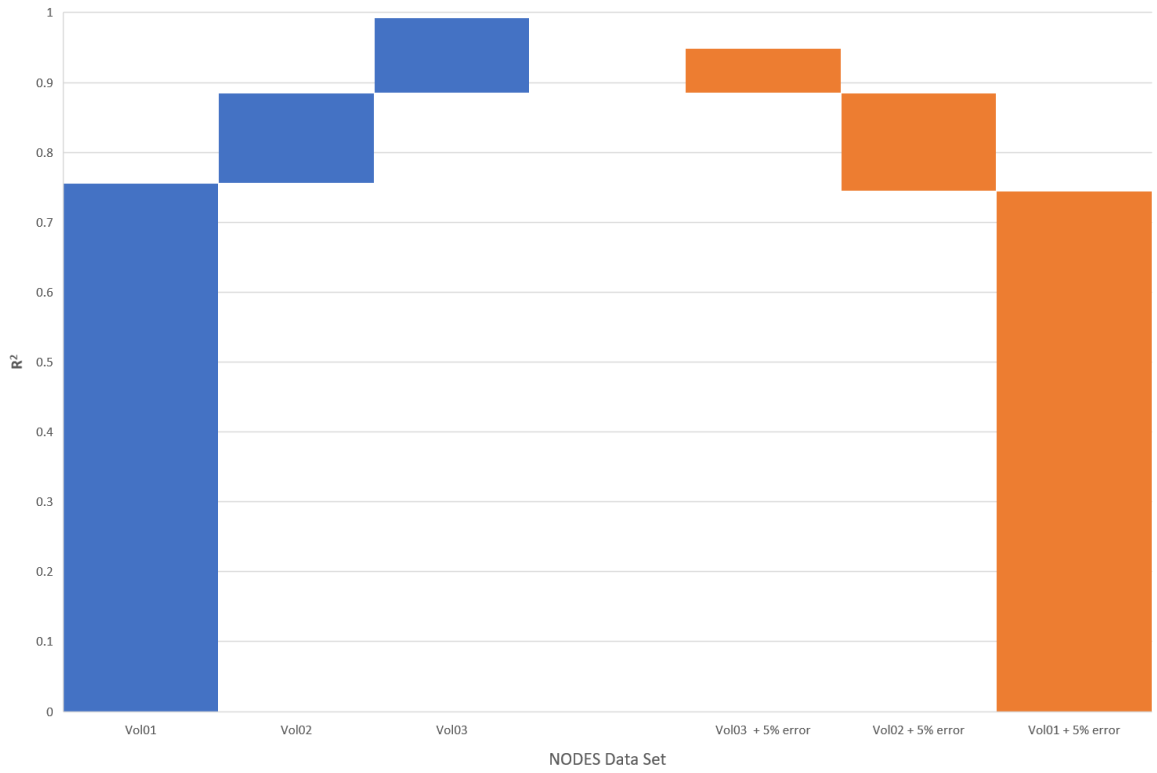


Figure 5: R^2 values for all NODES simulation datasets, comparing optimized results and the incremental improvement gained from additional solvent volumes (more R-slices). Blue: Error-free datasets. Orange: Datasets with 5% error applied

Given that Figure 3 already shows clear visual separation of good vs poor systems, these results indicate meaningful robustness in the CRYSS methodology.

Technical Approach

CRYSS employs a hybrid strategy combining:

- Monte Carlo simulation, which explores the parameter space broadly
- Focused optimization, which refines the best Monte Carlo candidates

This combination delivers excellent predictivity. Together, the results support an experimental strategy:

1) Run a large screen using only a single solvent volume per mixture.
2) Select the top ~25% of systems based on CRYSS-predicted R_{max}.
3) Perform more detailed experiments on this subset using two or more solvent volumes.

At its current build level, CRYSS is a GUI-free application. It:

1) accepts Excel input files containing composition data

2) parses replicates

3) clusters experiments by starting mixture composition

4) identifies R-slices (solvent volumes) via lever-rule analysis with error/outlier correction

5) solves for R_{max} accordingly

If a NODES dataset is provided, CRYSS plots R_{max}(NODES) vs R_{max}(CRYSS). For real experimental data (where true R_{max} is unknown), CRYSS outputs an Excel file containing mixture IDs and predicted R_{max} values.

In its present form, CRYSS is ideally suited for development into an API for integration with existing hardware/software platforms.

Expected Impact

CRYSS introduces a fundamentally new capability into crystallization development: the ability to convert small-scale equilibrium partitioning data into mechanistic predictions of maximum achievable yield (R_{max}) across multicomponent mixtures. This transforms early-stage purification screening from a largely empirical exercise into a quantitative, model-driven workflow.

By ranking solvent, solvent-mixture, and salt-formation systems based on their predicted purification ceiling, CRYSS enables process chemists to focus effort where it matters most. Instead of optimizing dozens of marginal systems, teams can rapidly triage large screening arrays and concentrate resources on the top-performing 20–25% of candidates. This reduces experimental burden, accelerates development timelines, and increases the likelihood of selecting a high-yielding crystallization route early in the project.

The hybrid Monte Carlo + optimization engine provides robustness against noise, allowing meaningful predictions even from minimal data (a single solvent volume per mixture). As a result, CRYSS supports a two-stage experimental strategy: broad, inexpensive initial screening followed by targeted, high-resolution follow-up studies. This approach is well aligned with modern parallel reactor platforms and high-throughput workflows.

Because CRYSS is mechanistic rather than empirical, it scales naturally to complex mixtures, salt-forming systems, and derivatization strategies. It also integrates seamlessly with NODES, enabling *in silico* exploration of purification landscapes before committing laboratory resources. In its current form, CRYSS is already suitable for API integration into automated screening hardware, ELNs, and laboratory informatics systems.

Taken together, CRYSS has the potential to reshape how crystallization screening is performed: from trial-and-error exploration to predictive, data-driven decision-making.

Conclusion

CRYSS V1.0 demonstrates that multicomponent crystallization behaviour — long considered too complex for practical mechanistic analysis — can be decoded, modelled, and used to guide purification strategy. The results from extensive NODES benchmarking show that even sparse equilibrium data contain enough information to estimate R_{max} with high fidelity, and that hybrid Monte Carlo + optimization methods deliver near-ideal predictivity when multiple solvent volumes are available.

The methodology is robust, scalable, and experimentally efficient. A single well-constructed equilibrium experiment can often provide a reliable estimate of R_{max}, and additional solvent volumes further refine the prediction. CRYSS therefore offers a practical route to mechanistic insight without imposing heavy experimental overhead.

Most importantly, CRYSS provides clarity at the earliest and most consequential stage of crystallization development: selecting which systems deserve attention. By identifying high-yielding solvent and salt-forming conditions before optimization begins, CRYSS helps ensure that development resources are spent on the right problems, not the wrong ones.

CRYSS V1.0 validates a concept that equilibrium partitioning data, interpreted correctly, can reveal the purification limits of complex mixtures. With this foundation established, future versions can expand into GUI interfaces, automated hardware integration and real-time analytics.

For now, CRYSS stands as a practical, mechanistic, and experimentally grounded tool that brings predictive power to crystallization screening — and opens the door to a more rational, data-driven future for purification science.

Acknowledgements

CRYSS would like to acknowledge the contribution made by the following people:

- Eric Damen
- Simone Darphorn-Hooijschuur
- Francois Gilardoni
- Ben McKay
- Lisa McQueen (née Agocs)
- Erik-Jan Ras
- Goran Verspui