NanoSpace Survey Report

Community priorities for next-generation astrochemical modelling

Mads-Peter V. Christiansen

2026-08-31

NanoSpace COST Action CA21126





This report is based upon work from COST Action NanoSpace, CA21126, supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

Executive summary

This consultation identifies priorities for next-generation astrochemical modelling within the NanoSpace community. The survey was open from 17 June to 14 August 2026 and received responses from 36 NanoSpace members. It covered input data, modelling components, current limitations, community priorities, and the potential role of machine learning (ML) and other data-driven methods.

Three findings stand out:

- **Needs are distributed rather than dominated by one bottleneck.** Respondents identified challenges across reaction and surface data, spectroscopy, chemical networks, validation, uncertainty information, and interoperability. Progress is therefore likely to require coordinated improvements across several connected areas.
- **Integration and validation are central concerns.** *Linking experiments, observations & models* was the leading individual priority, selected by 24 of 36 respondents. Benchmark datasets and validation cases also recurred across questions on priorities, modelling development, and barriers to ML.
- **ML has potential, but its use is conditional on stronger foundations.** Respondents identified applications including predicting missing rates and energies, surrogate modelling, calibration, and data extraction. Reliable training data, benchmark datasets, uncertainty estimates, and community standards were prominent constraints.

As an achievable near-term action, NanoSpace could coordinate a community-authored review of the current landscape of astrochemical databases and datasets, including their coverage, provenance, uncertainty information, interoperability, accessibility, and readiness for data-driven use. A maintained discovery catalogue could follow as a longer-term resource if the community can establish sustainable ownership and maintenance. Benchmark development, focused reaction-pathway and network activities, and targeted cross-disciplinary or ML support could provide practical test cases and complementary routes for collaboration.

These findings reflect a self-selecting sample of NanoSpace members. The planned direct consultations did not produce sufficient follow-up for a separate interview component, and the recommendations should therefore be understood as an analytical synthesis of the survey rather than evidence of formal community consensus.

1 Purpose and scope

This report summarises priorities identified by NanoSpace survey respondents concerning next-generation astrochemical modelling. The survey covered input data, modelling components, methodological limitations, and the potential role of machine learning (ML) and other data-driven methods.

2 Survey participation

The survey was distributed to NanoSpace Action members through several mailing-list invitations, and participation was voluntary. Responses were collected from 17 June to 14 August 2026, with 36 participants completing the survey. The first three questions concerned career stage, area of expertise, and systems or processes of interest; see Figure 3, Figure 1, and Figure 2.

Respondents represented a range of career stages, including 16 faculty members or professors. The largest areas of expertise were laboratory astrochemistry (9/36), computational chemistry and reaction modelling (7/36), observations and spectroscopy (7/36), and molecular simulations (6/36). Gas-phase chemistry was the most common system or process of interest (15/36), followed by spectroscopic data (9/36) and molecular clusters (5/36). Each remaining expertise and systems category was represented by one or two respondents.

Response counts vary between later questions. Unless stated otherwise, comparisons of rating distributions use substantive responses, excluding missing values and "Unable to judge" selections; the relevant denominators are therefore reported where they are central to the interpretation.

3 Priorities identified by respondents

Respondents selected up to three priorities from nine options, producing 107 selections; see Section 10.7 and Figure 7. *Linking experiments, observations & models* was the most frequently selected priority (24/36, approximately 67%), indicating that connections between empirical and modelling activities were a central concern among respondents.

Four explicitly data-focused options accounted for 48 of the 107 selections:

- Better reaction-rate data (14/36),
- Benchmark datasets & validation cases (13/36),
- Better binding/desorption energy data (11/36),
- ML-ready databases (10/36).

This combined total is a post-hoc analytical grouping rather than a category presented in the survey. It nevertheless suggests that data availability, quality, validation, and usability were recurring concerns. Modelling-related priorities also received attention, particularly *Improved chemical networks* (14/36) and *More realistic surface/ice models* (9/36).

4 Current data and modelling limitations

The survey examined the importance of different forms of input data, the extent to which missing or uncertain data are limiting, and the need for further development of selected modelling components. Across the eight input-data categories, 90–97% of substantive responses rated each category as "Important," "Very important," or "Critical"; see Figure 4. Spectro-

scopic reference data, binding and desorption energies, and surface-reaction rates received particularly high numbers of "Very important" and "Critical" ratings. The similarity of the distributions supports a broad set of complementary inputs rather than a definitive ranking among them.

For uncertainties and missing data, 73–86% of substantive responses rated every listed area as at least "Limiting"; see Figure 5. Although "Limiting" was often the most common rating, no area emerged as an unambiguous principal constraint. The responses therefore suggest that perceived limitations are distributed across several types of data rather than concentrated in a single deficiency.

Ratings of modelling components showed somewhat greater differentiation; see Figure 6. Reaction-pathway modelling had the strongest indication of further development needs, with 24 of 31 substantive responses rating the need as "Major" or "Critical." Grain and ice-surface models (18/29), uncertainty propagation (17/27), and benchmark datasets (16/28) also received substantial numbers of these ratings. Comparatively many respondents selected "Unable to judge" for kinetic Monte Carlo (12/36) and interatomic potentials and force fields (9/36), limiting direct comparisons for these more specialised components.

5 Perspectives on machine learning

The survey included two questions on ML and data-driven methods: one on potential applications and one on barriers to their use. For potential applications, 35 respondents made 97 selections, comprising 96 selections from the predefined options and one write-in response; see Figure 8. The leading applications were *Predicting missing reaction rates* (16), *Predicting binding energies* (15), *Surrogate models for expensive simulations* (14), *Model calibration against observations* (13), *Literature/data extraction* (12), and *Interatomic potentials* (11). *Automated reaction-network construction* and *Uncertainty quantification* each received six selections.

Of the 35 respondents, 32 selected at least one potential application, while three selected only *No clear role at present*. The emphasis on predicting missing rates and energies may connect the perceived potential of ML to the data needs identified elsewhere in the survey. Similarly, interest in model calibration may reflect the broader priority placed on connecting models with empirical evidence. One respondent also proposed using data-driven methods to interpret and integrate findings across disciplines.

The question on barriers received 90 selections from 34 respondents; see Figure 9. The leading barriers were *Lack of reliable training data* (20), *Lack of benchmark datasets* (16), *Poor uncertainty estimates* (12), and *Lack of community standards* (11). The first two reinforce the prominence of data quality and validation in the priority results. Poor uncertainty estimates may also make it difficult to determine when ML predictions are reliable.

The meaning of "community standards" was not specified in the survey. The thematic summary of the open-ended responses includes differences in terminology and notation across overlapping disciplines; see Section 10.10. This may be related to the reported concern about standards, although respondents may also have had dataset quality, curation, provenance, or other forms of standardisation in mind.

6 Limitations of the consultation

The results should be interpreted in light of several limitations. Participation was voluntary, and the 36 respondents formed a self-selecting sample of NanoSpace members rather than a representative sample of the entire community. The distribution of expertise among respondents and the number of substantive responses also varied between questions. Missing responses and selections of "Unable to judge" therefore limit comparisons, particularly for specialised modelling components.

The questions on priorities, potential applications of ML, and barriers to its use allowed respondents to select multiple options. Their results indicate recurring interests among respondents, but they do not constitute a strict ranking or a test of statistical significance. The qualitative evidence is more limited: 13 respondents answered the open-ended question on underappreciated data gaps or modelling limitations, while nine answered the question on relevant datasets, tools, models, or community resources. These responses provide examples and context but should not be interpreted as evidence of how prevalent a concern is across NanoSpace.

The original plan also included direct consultations with selected experts. Outreach was made during the survey period, but the level of follow-up was insufficient to support a separate consultation component. Consequently, this report should be understood as a synthesis of the survey rather than as a combined analysis of survey responses and expert interviews.

An anonymised workbook containing responses to Questions 1–9 has been provided separately to the Action for internal review and should undergo an additional disclosure review before any unrestricted public release.

7 Synthesis of findings

7.1 A distributed data problem

The responses do not identify a single data type or modelling component as the dominant bottleneck. Instead, they indicate needs across reaction rates, binding and desorption energies, spectroscopy, surface processes, chemical networks, and uncertainty estimates. The open-ended responses illustrate the breadth of these concerns, mentioning uncertain or unknown reactions, limited spectroscopic information in particular regimes, sparse benchmark coverage, and the use of fixed parameters to represent heterogeneous grain and ice surfaces. Because these points were raised by individual respondents, they are best treated as examples of possible gaps rather than as independently established community priorities.

7.2 Integration and validation

The clearest individual priority was *Linking experiments, observations & models*, selected by 24 of 36 respondents. The attention given to benchmark datasets and validation cases reinforces the importance of making results from different parts of the community easier to compare and use together. The open-ended responses also included differences in terminology and notation across overlapping disciplines. Taken together, these results suggest that coordination, validation, and interoperability are important alongside the generation of additional data.

7.3 Conditional potential for machine learning

Most respondents to the question on ML selected at least one potential application, particularly the prediction of missing reaction rates and binding energies and the construction of surrogate models for expensive simulations. At the same time, the most frequently selected barriers were a lack of reliable training data and benchmark datasets. Poor uncertainty estimates and a lack of community standards were also recurring concerns. The survey therefore indicates interest in data-driven methods, but it also suggests that their responsible use depends on improvements in data quality, documentation, validation, and uncertainty reporting.

8 Recommendations and next steps

The following recommendations offer a staged pathway: first mapping existing resources and gaps, then testing integration through bounded scientific activities, continuing targeted support for skills and cross-disciplinary collaboration, and finally considering longer-term discovery infrastructure. They are evidence-informed options rather than a formal ranking or evidence of community consensus.

8.1 Develop a community-authored landscape review

NanoSpace could initiate a cross-disciplinary review paper mapping the current landscape of databases and datasets relevant to astrochemical modelling. Before prioritising the production of additional data, such a review could establish which resources already exist, what they cover, and how readily they can be discovered and reused.

The review could cover the major data types, physical systems, processes, and properties required by astrochemical models. For each resource, it could consider scope and coverage, access interfaces and formats, metadata, uncertainty information, provenance, and accessibility. It could also assess interoperability, alignment with FAIR data principles, and readiness for data-driven or ML applications. This would provide a basis for identifying gaps, overlaps, and barriers to combining resources across subfields.

Respondents named resources spanning chemical networks and kinetics, spectroscopy and photoprocesses, surface data, and optical-property modelling; see Section 10.11. These suggestions are examples of resources that a landscape review could examine; they are not a complete inventory and were not evaluated as part of this survey. Producing a useful review would require contributions from observational, laboratory, theoretical, modelling, and data-driven communities.

8.2 Create a benchmark-and-validation pilot

NanoSpace could develop one or two bounded reference cases that connect experimental measurements, observations, and model predictions. Each case could specify reference inputs and outputs, comparison metrics, uncertainty information, provenance, and versioning. Such a pilot would test in practice what is required to make results from different parts of the community comparable and reusable.

This recommendation is supported by converging results from several survey questions. *Linking experiments, observations & models* was selected as a priority by 24/36 respondents, while *Benchmark datasets & validation cases* was selected by 13/36. Further development of benchmark datasets was rated “Major” or “Critical” by 16/28 substantive respondents, and

16/34 respondents selected a lack of benchmark datasets as a barrier to ML and data-driven methods. This recurrence makes a benchmark-and-validation pilot a particularly well-supported option, although the survey does not identify which scientific cases it should address.

8.3 Launch a focused reaction-pathway and network-development activity

NanoSpace could convene a focused activity to examine a scientifically important family of reactions or pathways. A scoped gap analysis could identify missing reactions, uncertain rates, inconsistencies among chemical networks, and high-impact targets for laboratory measurements or theoretical calculations. Its outputs could also provide a concrete test case for the benchmark-and-validation pilot.

Reaction-pathway modelling had the clearest development signal in the survey: 24/31 substantive respondents rated its need for further development as "Major" or "Critical." *Improved chemical networks* and *Better reaction-rate data* were each selected as priorities by 14/36 respondents, and 16/35 respondents selected predicting missing reaction rates as a potential application of ML or data-driven methods. These results support a focused activity, but they do not determine its chemical or astrophysical scope; that choice would require further scientific consultation.

8.4 Continue targeted cross-disciplinary and ML support

NanoSpace has already invested in this area through the August 2025 NanoSpace AI in Astrochemistry Training School, which combined foundations in ML with lectures and practical exercises on applications in chemistry and astrochemistry. Further support could build on that foundation through targeted workshops, paired projects, or working groups organised around concrete datasets, benchmarks, and modelling problems, rather than treating training as a stand-alone activity.

The survey indicates that such support remains worthwhile. Limited access to ML expertise was selected as a barrier by 10/34 respondents, while *Linking experiments, observations & models* was selected as a priority by 24/36 and model calibration against observations as a potential ML application by 13/35. The survey did not evaluate the 2025 training school or directly ask whether another school should be held. It nevertheless suggests a continuing role for activities that connect methodological expertise to specific cross-disciplinary research problems.

8.5 Consider a discovery-oriented catalogue

As a possible later stage, NanoSpace could assess the feasibility of maintaining a curated catalogue that helps researchers discover relevant datasets. Its scope and metadata could be informed by the landscape review and by practical requirements identified through the benchmark and reaction-pathway activities. The catalogue would not own, host, or duplicate the underlying data. Instead, it would describe resources and link users to the authoritative dataset owners.

Entries could be searchable or filterable by system, process, property, and data type. Each entry could provide concise information on coverage, format, uncertainty, provenance, access conditions, and licensing, together with a link to the source. A transparent process would also be needed for proposing, reviewing, updating, and retiring entries.

The principal challenge would be maintenance rather than initial implementation. The catalogue should therefore proceed only if the community can identify named ownership, an appropriate review process, and a realistic update schedule. A small pilot based on the scope developed in the landscape review could test its usefulness and maintenance requirements before NanoSpace commits to a larger resource.

9 Conclusion

The survey points to a need for coordinated improvements across input data, modelling components, validation practices, uncertainty information, and interoperability. It does not identify a single deficiency whose resolution would be sufficient on its own. Instead, the results emphasise connections between experiments, observations, and models, supported by data and benchmarks that can be assessed and reused across disciplinary boundaries.

A community-authored landscape review offers a practical near-term step towards this goal. A benchmark-and-validation pilot is especially well supported because related needs recur across the questions on priorities, modelling development, and barriers to ML. A focused reaction-pathway and network activity could provide one source of scientifically grounded test cases, while continued cross-disciplinary and ML support could help participants apply appropriate methods to concrete data and modelling problems. A maintained discovery catalogue could build on the landscape review and lessons from these activities as a longer-term resource, provided that NanoSpace first establishes sustainable ownership and maintenance. These recommendations are a synthesis of the survey findings rather than evidence of a formal community consensus or a statistically established ranking of interventions.

10 Appendix

The appendix includes figures and qualitative summaries based on the survey responses.

For the open-ended questions, responses were reviewed and grouped thematically. Blank, uncertain, and non-substantive entries were excluded from thematic interpretation but retained in the response totals. Substantive answers were paraphrased, and distinctive details were generalised where necessary to reduce the risk of identifying individual respondents. A single response could inform more than one theme. The themes are illustrative: their order does not constitute a ranking, and they should not be interpreted as estimates of prevalence across NanoSpace.

10.1 Question 1

What area best describes your main expertise?

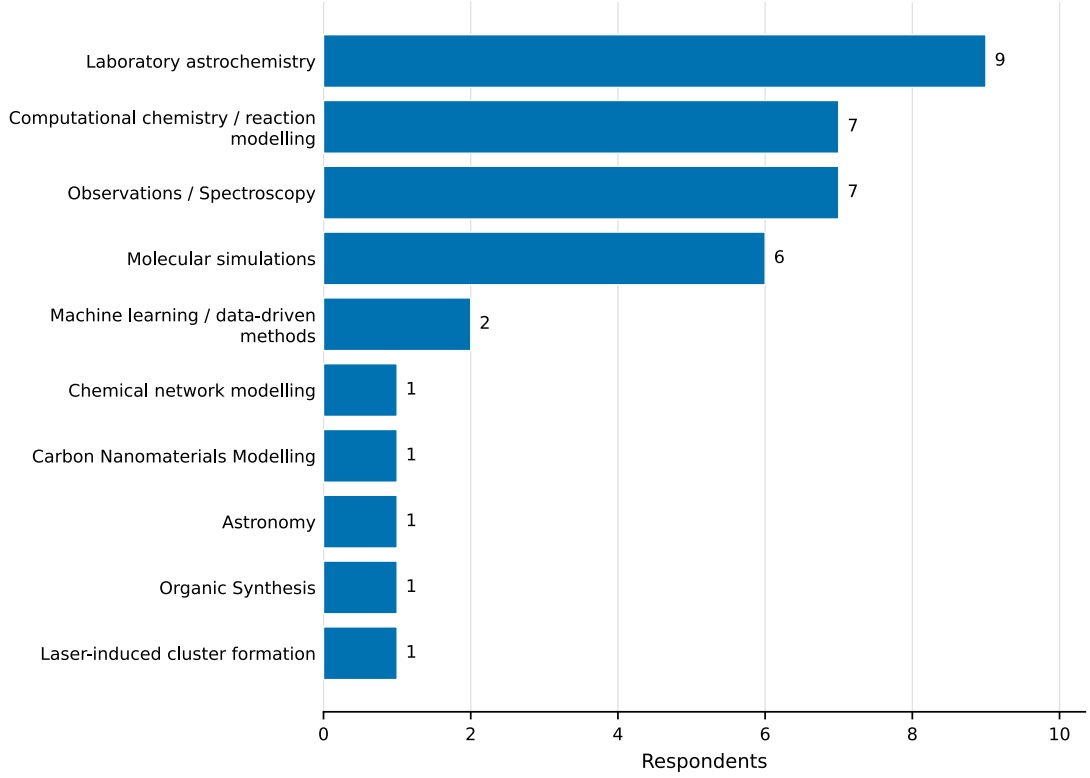


Figure 1: Main areas of expertise among respondents (n = 36).

10.2 Question 2

What type of systems or processes do you mainly work on?

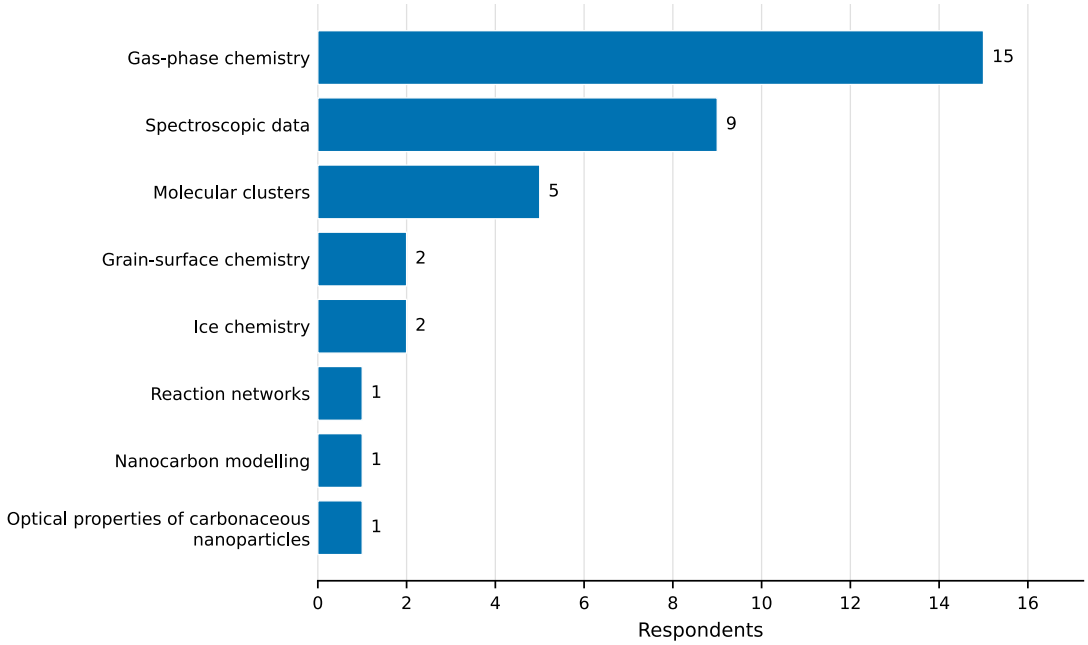


Figure 2: Main systems or processes studied by respondents (n = 36).

10.3 Question 3

What career stage are you at?

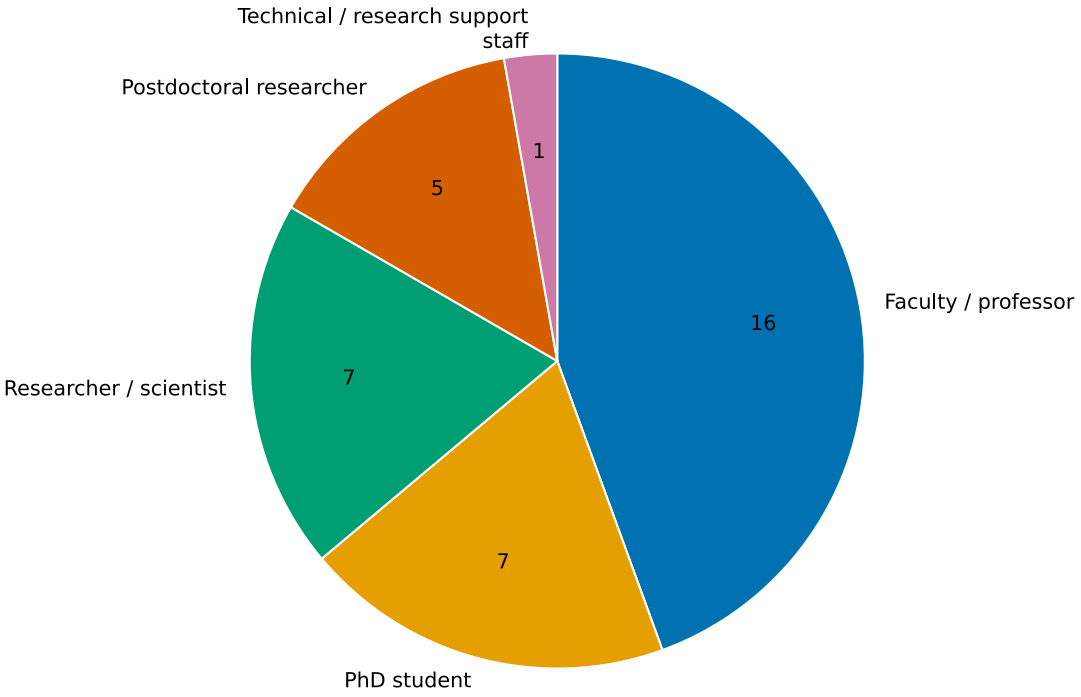


Figure 3: Career stages of respondents (n = 36).

10.4 Question 4

How important are the following *input data* for next-generation astrochemical models?

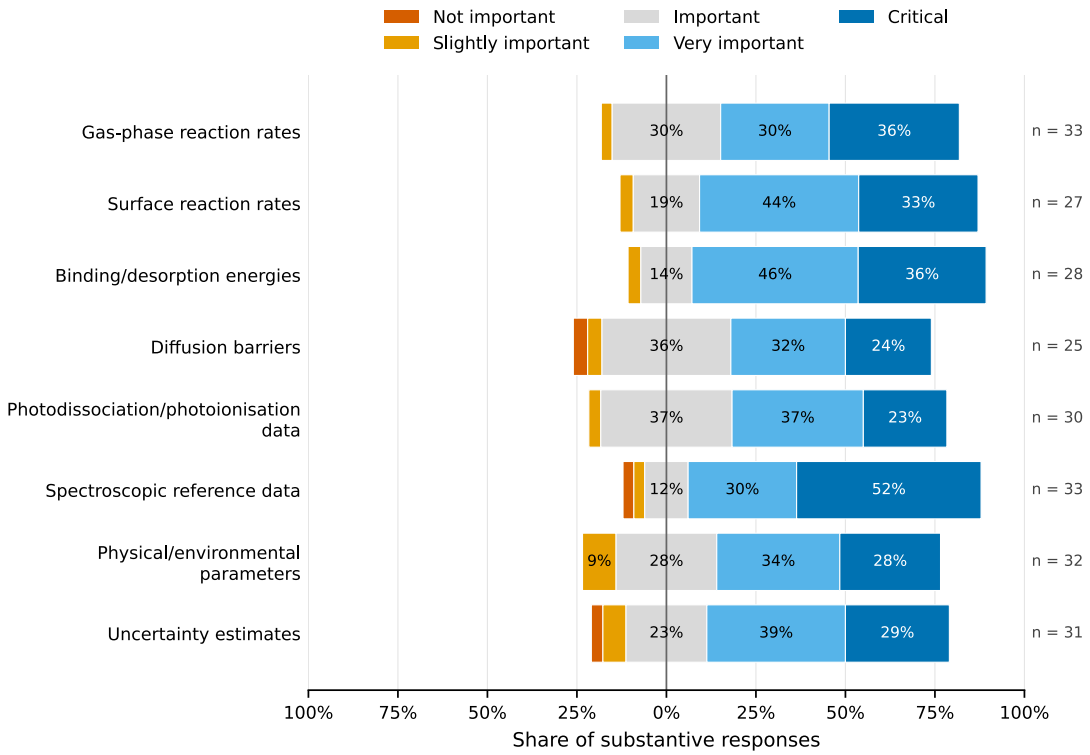


Figure 4: Distribution of substantive importance ratings for different types of **input** data. Bars show percentages and right-hand labels report the substantive sample size; "Unable to judge" and missing responses are excluded.

10.5 Question 5

How limiting are current uncertainties or missing data in the following areas?

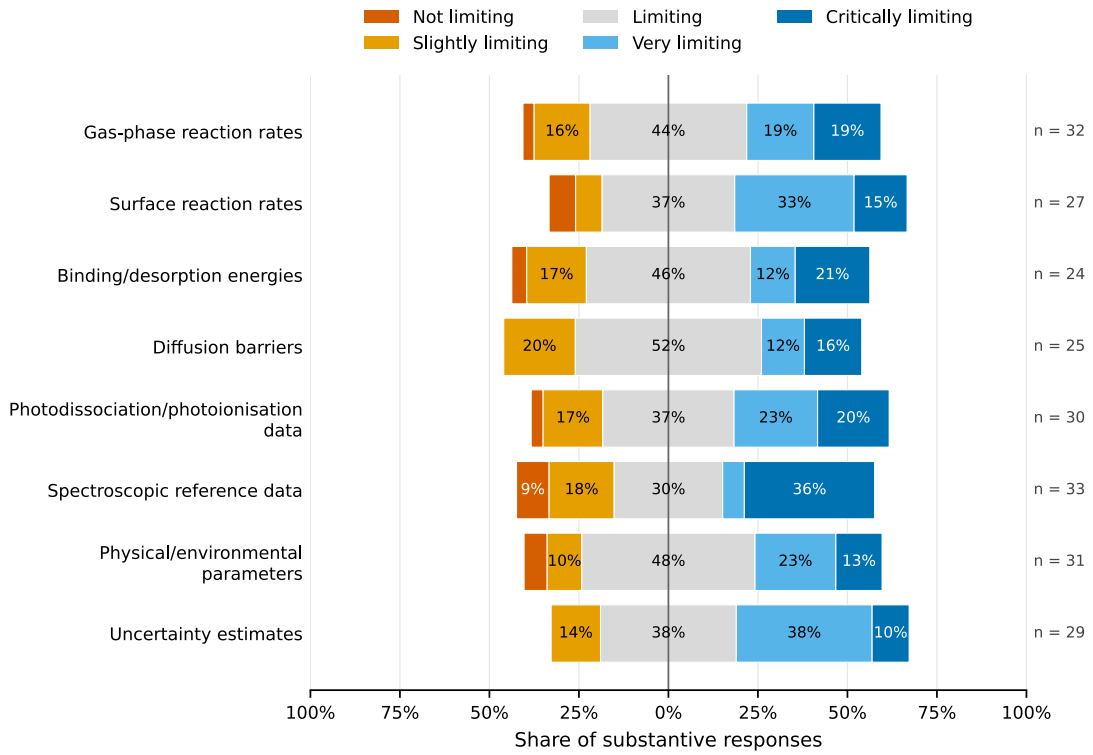


Figure 5: Distribution of substantive ratings of limitations arising from uncertainties or missing data. Bars show percentages and right-hand labels report the substantive sample size; "Unable to judge" and missing responses are excluded.

10.6 Question 6

How much further development is needed in the following modelling components to support next-generation astrochemical modelling?

Please rate how much further development is needed for each modelling component. "Adequate" means that current approaches are broadly sufficient, while "Critical" means that lack of development is a major bottleneck for progress.

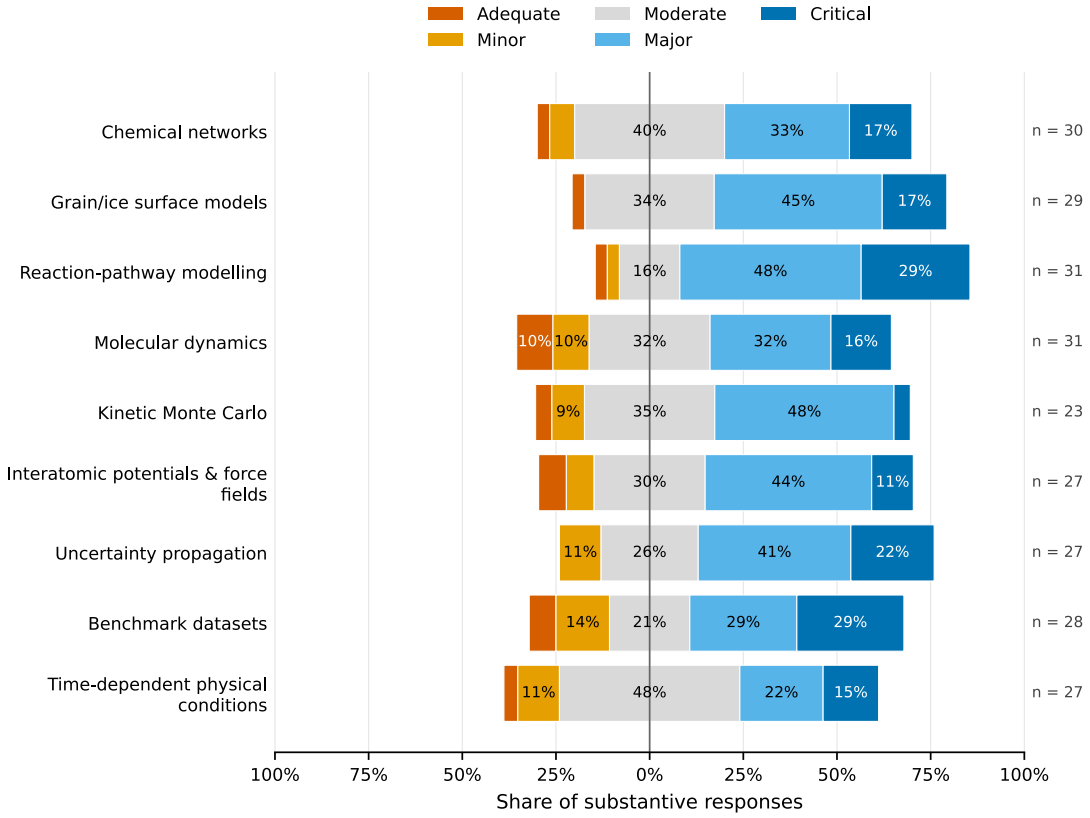


Figure 6: Distribution of substantive ratings of the need for further development. Bars show percentages and right-hand labels report the substantive sample size; "Unable to judge" and missing responses are excluded.

10.7 Question 7

From the list below, select the three most important priorities for the NanoSpace community to address.

For this question, participants could choose up to three options, including a free-form answer.

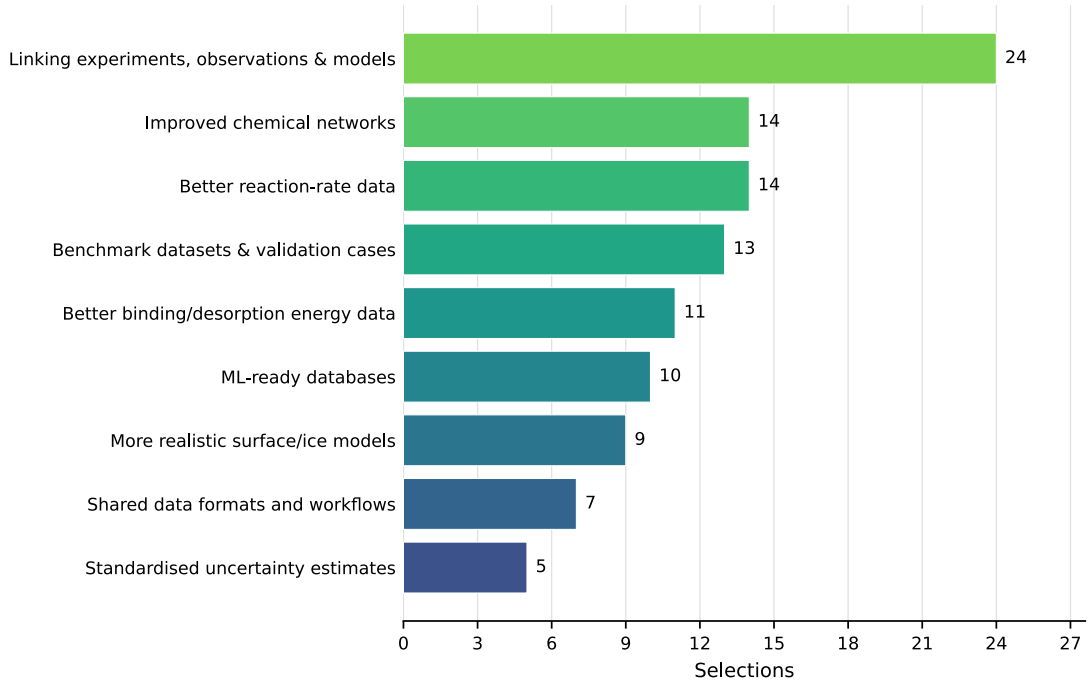


Figure 7: Priorities selected by respondents (107 selections from 36 respondents).

10.8 Question 8

Where do you see the greatest potential for machine learning or data-driven methods?

For this question, participants could choose up to three options, including a free-form answer.

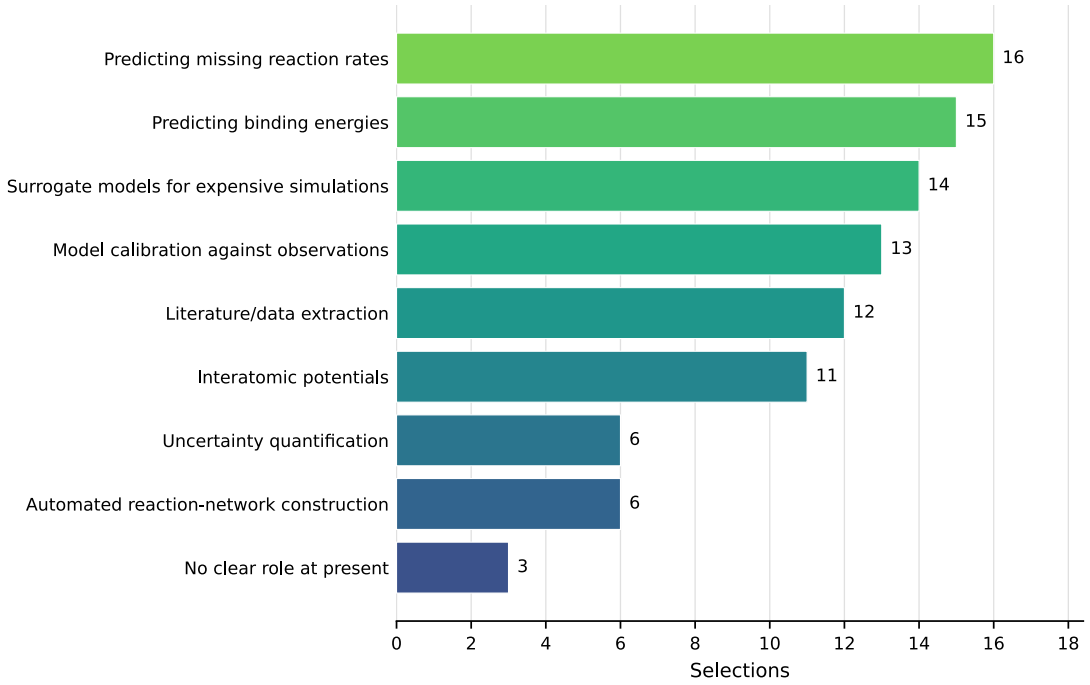


Figure 8: Potential applications of machine learning and data-driven methods (96 predefined selections from 35 respondents, plus one write-in response).

One respondent additionally suggested using data-driven methods to support the interpretation and integration of research findings across disciplines.

10.9 Question 9

What are the main barriers to using ML/data-driven methods in this area?

For this question, participants could choose up to three options, including a free-form answer.

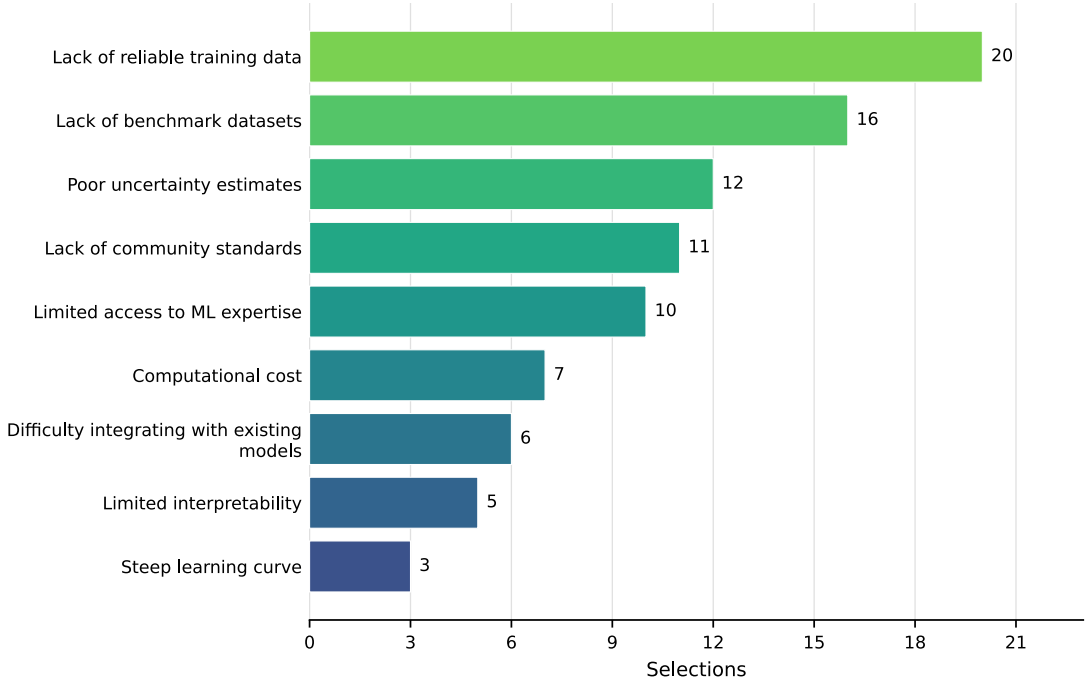


Figure 9: Main barriers to the use of ML and data-driven methods (90 selections from 34 respondents).

10.10 Question 10

What is one data gap or modelling limitation that you think is currently underappreciated?

Thirteen respondents supplied an entry. Nine entries identified a sufficiently clear data gap or modelling limitation for thematic interpretation; four were uncertain, non-substantive, or did not identify a specific gap. The substantive responses were organised into the following unranked themes:

- **Incomplete or uncertain physical and chemical data.** Respondents described missing reactions and uncertain rates, energies, diffusion barriers, and surface parameters. They also questioned the use of fixed values for grain and ice properties that vary with physical conditions.
- **Spectroscopy, excitation, and radiative processes.** Responses identified gaps in spectroscopic information for transient species and some wavelength regimes, alongside challenges involving excitation, radiative transfer, fluorescence, and unimolecular processes.
- **Benchmarking and model validation.** Respondents noted sparse benchmark coverage and concern that models with many adjustable inputs may reproduce observations without uniquely validating the underlying physics.
- **Terminology and interdisciplinary interoperability.** Responses identified differences in notation, terminology, and conventions across disciplines as barriers to interoperability.

10.11 Question 11

Are there specific datasets, tools, models, or community resources that should be considered in this report?

Nine respondents supplied an entry. Five entries named a resource or provided a substantive resource-related suggestion; four did not identify a specific resource. The public resources named by respondents are consolidated below.

Area	Respondent-named examples
Chemical networks and kinetics	KIDA; UMIST/UDfA; UCLCHEM
Spectroscopy and photoprocesses	CDMS; JPL molecular spectroscopy catalogue; HITRAN; NASA Ames PAH IR Spectroscopic Database; Leiden photodissociation and photoionisation database
Surface and binding-energy data	BEEP
Optical-property modelling	DADI

Respondents also suggested integrating machine learning with established modelling approaches and developing benchmarks connecting calculations, simulations, measurements, and astrochemical models. Desired features included standardised metadata, uncertainty information, and formats suitable for FAIR and data-driven use. These suggestions were not independently evaluated and are not exhaustive.