Faraday Discussions
Cite this: *Faraday Discuss.*, 2025, **256**, 373

DISCUSSIONS

View Article Online
View Journal | View Issue

Discovering structure–property correlations: general discussion

Andy S. Anker, Alán Aspuru-Guzik, Chiheb Ben Mahmoud, Sophie Bennett, Ksenia R. Briling, Arya Changiarath, Sanggyu Chong, Christopher M. Collins, Andrew I. Cooper, Daniel Crusius, Kevion K. Darmawan, Basita Das, Nicholas David, Graeme M. Day, Volker L. Deringer, Fernanda Duarte, Annabel Eardley-Brunt, Matthew L. Evans, Rob Evans, Ian Fairlamb, Barnabas A. Franklin, Jeremy Frey, Alex M. Ganose, Mark Goulding, Roohollah Hafizi, Matthijs Hakkennes, Niamh Hickey, Gillian James, Kim E. Jelfs, Adarsh V. Kalikadien, Venkat Kapil, Zsuzsanna Koczor-Benda, Ferdinand Krammer, Heather J. Kulik, Vishank Kumar, Christian Kuttner, Erwin Lam, Yuchen Lou, Eltjo Mante, Jennie Martin, Austin M. Mroz, Tahereh Nematiaram, Charles W. P. Pare, Sarbani Patra, James Proudfoot, Branko Ruscic, Matthew R. Ryder, Ken Sakaushi, Jörg Saßmannshausen, Brett M. Savoie, Nadine Schneider, Philippe Schwaller, Bastian Bjerkem Skjelstad, Wenhao Sun, Filip T. Szczypiński, Steven Torrisi, Katharina Ueltzen, Shubham Vishnoi, Aron Walsh, Xinwei Wang, Chloe Wilson, Ruiqi Wu and Jakob Zeitler

DOI: 10.1039/d4fd90062f

Jakob Zeitler opened the discussion of the paper by Kim E. Jelfs: What is the competitive strategy for using open source?

While open source suggests open and transparent building of software, in reality open source has established itself as a commercial strategy for corporations to strategically break competitors' advantages or dominate adjacent markets.

A key challenge in competing with open source funded by big corporations is the budget size. Corporations can deploy multiple full-time positions over many years to advance their open source projects. Academia simply cannot, especially with their regular turnover.

Therefore, unless the adoption of academia-driven open source projects happens fast and attracts other talents and contributors, no original maintainers will be left to continue the project, which would be a loss.

Therefore, it is crucial to decide on a strategy for long-term maintenance and the 'competitive' USP to be provided through that to make an academia-driven open source project worthwhile.

Kim E. Jelfs replied: Code and software maintenance is an ongoing challenge in the academic community because of the high turnover rate and the limited funding calls prioritizing software development and maintenance in science. While a challenge, this landscape is changing with the UKRI and EPSRC announcing funding for research software engineers, which is critical to maintaining, supporting and developing projects like Web-BO.

Steven Torrisi said: A remark that may be relevant particularly to students and postdocs in the room is that, coming from the industry perspective, I'd like to strongly concur with your premise that making these tools more accessible to users really matters for real-world impact. This kind of method is still cutting-edge, and tools like this that 'reduce it to practice' are critical. Don't underestimate the extent to which, when you interact with people with traditional training and background in the physical sciences and engineering, knowing how all of this works is still a professional superpower and holds value for them – I particularly like some of the decisions you've made for, *e.g.*, removing complexity for the user. To that end, my question is about the end user. Have you had the chance to put this in the hands of a traditional chemist to see what their stumbling blocks are, and if so, what did you learn?

Kim E. Jelfs answered: We are just beginning to do this, and for the moment it has been very helpful to remove software bugs and to make the app clearer to use! As the app is used more and we see what experimental chemists find to be the stumbling blocks, we will share this in future updates.

Alán Aspuru-Guzik addressed Kim E. Jelfs and Austin M. Mroz: Your paper (**https://doi.org/10.1039/d4fd00109e**) is a great and beautiful effort to popularize Bayesian Optimization (BO) for non-experts and experts alike! As a platform that intends to be widely used, I wonder if you have considered extending the platform to include more models and benchmarking tools. I am particularly thinking about the recently-released models by BASF[1] as well as Pfizer, as well as my own group's Atlas.[2] Another benchmarking platform to consider is Olympus.[3,4] This would allow the user to access more capabilities as well as compare algorithms.

1 J. P. Dürholt, T. S. Asche, J. Kleinekorte, G. Mancino-Ball, B. Schiller, S. Sung, J. Keupp, A. Osburg, T. Boyne, R. Misener, R. Eldred, W. S. Costa, C. Kappatou, R. M. Lee, D. Linzner, D. Walz, N. Wulkow and B. Shafei, BoFire: Bayesian Optimization Framework Intended for Real Experiments, *arXiv*, 2024, preprint, arXiv:2408.05040v1 [cs.LG], DOI: **10.48550/arXiv.2408.05040**.

2 R. Hickman, M. Sim, S. Pablo-García, I. Woolhouse, H. Hao, Z. Bao, P. Bannigan, C. Allen, M. Aldeghi and A. Aspuru-Guzik, Atlas: A Brain for Self-driving Laboratories, *ChemRxiv*, 2023, preprint, DOI: **10.26434/chemrxiv-2023-8nrxx**.

3 F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, M. Christensen, E. Liles, J. E. Hein and A. Aspuru-Guzik, Olympus: a benchmarking framework for noisy optimization and experiment planning, *Mach. Learn.: Sci. Technol.*, 2021, **2**, 035021, DOI: **10.1088/2632-2153/abedc8**.

4 R. Hickman, P. Parakh, A. Cheng, Q. Ai, J. Schrier, M. Aldeghi and A. Aspuru-Guzik, Olympus, enhanced: benchmarking mixed-parameter and multi-objective optimization in chemistry and materials science, *ChemRxiv*, 2023, preprint, DOI: **10.26434/chemrxiv-2023-74w8d**.

Austin M. Mroz responded: Thank you for your kind sentiments. Web-BO is in the initial development phases and we are actively working towards integrating more complex algorithms into the back end. This is definitely a priority, as most chemical tasks do not fit within the single-fidelity-single-objective regime that is currently released. Indeed, BO for chemical applications is an incredibly rich field and we hope to represent that within Web-BO by including state-of-the-art algorithms released by BASF, Pfizer, your group, and others.

Sarbani Patra asked: To what extent can the Web-Bo front end be considered a black box? How deeply must one understand the mathematical aspects of the Bayesian optimization process to use Web-Bo? Is it possible to simply supply reasonable parameters for optimization and expect useful results for guiding experiments without being familiar with the internal machinery of the optimization process? In other words, how much knowledge about Bayesian optimization would be deemed sufficient to obtain useful insights from Web-Bo?

Kim E. Jelfs answered: You could indeed use it as black box, but of course this is not desirable in general and would be likely to reduce the performance of the Bayesian optimisation for many users. This is why we are approaching providing training to experimental researchers to use this (and this will be followed by further documentation online).

Charles W. P. Pare addressed Kim E. Jelfs and Austin M. Mroz: Looking at the nature of your target audience (academic/industrial researchers), showing how the BO made the decision could be of interest in building trust in the algorithm, and in the interest of understanding the relationships that govern the design space, users, especially researchers, could benefit from variable effects on the response under investigation.

Austin M. Mroz answered: We agree; lack of trust in Bayesian optimization on the part of the researchers is one of the barriers to realizing the full power of BO for chemical tasks. This is an important topic for our future work and we are currently exploring different ways of addressing this challenge.

Matthew L. Evans remarked: As one of the developers of datalab, we still find it difficult to demonstrate the benefits of data-sharing on the level of individual researchers. It's great to see tools like Web-BO integrating directly with data platforms/ELNs, which will certainly help in this regard. Do you see a future where information from multiple datalab (or otherwise) instances from different labs could be combined in a federated manner and used by tools like Web-BO to provide even better suggestions or prediction? Do we need to better define the hypotheses that our data is trying to address ahead of time to make this possible?

Kim E. Jelfs responded: This is obviously a really attractive scenario and of course we benefit from the digital chemistry field in general having a positive attitude to open-sourcing data and software. We think it is very plausible for collaborators to make use of datalab combined with Web-BO to have federated data to accelerate discovery. It remains challenging to do that on a larger scale, but hopefully early examples can demonstrate the benefit to others. We need to also discuss further how the data generators can be properly recognised.

Alán Aspuru-Guzik added: Thank you for releasing datalab open source and also for maintaining it as well as all the other projects you maintain. We indeed developed a multi-lab solution called ORGANIZE-IT that we used for our Organic Laser project.[1] We also use ioChem-BD (**https://www.iochem-bd.org/**) and AiiDA (**https://www.aiida.net/**) as part of ChemOS 2.0.[2] I believe that standards will arise and that many of us will start sharing data amongst labs as the revolution of self-driving laboratories continues to take over.[3]

1 F. Strieth-Kalthoff, H. Hao, V. Rathore, J. Derasp, T. Gaudin, N. H. Angello, M. Seifrid, E. Trushina, M. Guy, J. L. Liu, X. Tang, M. Mamada, W. Wang, T. Tsagaantsooj, C. Lavigne, R. Pollice, T. C. Wu, K. Hotta, L. Bodo, S. Y. Li, M. Haddadnia, A. Wolos, R. Roszak, C. T. Ser, C. Bozal-Ginesta, R. J. Hickman, J. Vestfrid, A. Aguilar-Granda, E. L. Klimareva, R. C. Sigerson, W. D. Hou, D. Gahler, S. Lach, A. Warzybok, O. Borodin, S. Rohrbach, B. Sanchez-Lengeling, C. Adachi, B. A. Grzybowski, L. Cronin, J. E. Hein, M. D. Burke and A. Aspuru-Guzik, Delocalized, asynchronous, closed-loop discovery of organic laser emitters, *Science*, 2024, **384**, eadk9227, DOI: **10.1126/science.adk9227**.

2 M. Sim, M. G. Vakili, F. Strieth-Kalthoff, H. Hao, R. J. Hickman, S. Miret, S. Pablo-García and A. Aspuru-Guzik, ChemOS 2.0: An orchestration architecture for chemical self-driving laboratories, *Matter*, 2024, **7**, 2959–2977, DOI: **10.1016/j.matt.2024.04.022**.

3 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid and A. Aspuru-Guzik, Self-Driving Laboratories for Chemistry and Materials Science, *Chem. Rev.*, 2024, **124**, 9633–9732, DOI: **10.1021/acs.chemrev.4c00055**.

Jennie Martin asked: Is it possible within Web-BO or Bayesian optimization processes in general to change the optimization space mid-way through the process; for instance, if it becomes apparent that a given variable does not impact the target?

Kim E. Jelfs answered: Yes, this is possible in the sense that you would define a new optimization "campaign". Indeed, paying close attention to the optimization trajectory is critical to ensuring positive outcomes. If it becomes apparent that a given variable doesn't impact the target, this may be removed; however, you would start from the beginning, with more initial datapoints.

Andy S. Anker remarked: Thank you for making Web-BO an open-source project! I have noticed that much of the validation for Bayesian optimisation algorithms, as well as other machine learning models, tends to rely on benchmarks derived from simulated data. You mentioned that the Web-BO project is intended to be implemented across the entire department, placing you in a unique position where it will be applied to a variety of experimental tasks in chemistry. Do you plan to open-source these experimental data benchmarks in the future, to enable more robust validation of Bayesian optimisation algorithms?

Kim E. Jelfs answered: We would certainly anticipate the experimental data from these projects to be shared alongside publication. We also anticipate in the future writing a reflective article on the experiences across different chemical areas, which we are indeed well placed to do across a department.

Philippe Schwaller addressed Austin M. Mroz and Kim E. Jelfs: What do you communicate to your experimental partners? How do you valorise the experiments they have spent months on?

Austin M. Mroz replied: Within the context of Web-BO, we aim to put the power of Bayesian optimization (BO) into the hands of the experimentalists. Thus, our communication is largely comprised of training for the platform and a general understanding of BO. Throughout this process, we emphasize that not all optimization trajectories are fruitful and that problem formulation is one of the most significant determinants when it comes to optimization performance.

Indeed, BO truly thrives when the objective function is unknown and evaluations are expensive—within chemistry, this can be because of resource expenditure, time, *etc.* This also puts BO at a slight disadvantage; when more resources are required to perform an experiment (evaluate the objective function), it takes more trust in the BO algorithm for the researcher to expend those resources. Indeed, within BO it is sometimes unclear why certain experiments are suggested and may engender distrust in the researcher. Thus, building trust is essential to realizing the full advantages of BO in chemistry and is definitely an outstanding challenge to address.

Erwin Lam queried: How do you see the next steps to engage/promote the community from the experimental side to use Web-BO (*e.g.* performing a case study)?

Kim E. Jelfs responded: We have recently run a workshop with about 25 participants to provide training in Bayesian optimisation in general and also Web-BO, so we anticipate some examples from this. We are also beginning projects with several groups across different chemical topics that will provide case studies.

Jakob Zeitler asked: How do you reuse the data if you change the input?

It was stated that Web-BO can flexibly adjust the campaign inputs during the campaign in two ways:

(i) Adding new parameters.

(ii) Removing parameters.

For case (ii), we can remove parameters, but for the data collected so far to be reused in the surrogate model, we need to ensure that experiments going ahead fix the value of the removed parameter.

For case (i), adding parameters is only possible if we have data on what parameter value they were used in previous samples. If not, we are introducing confusion, which is detrimental to the convergence process and will increase the noise that needs to be modelled.

Therefore, it is undesirable to allow users to change the input parameter space during an experimental run. Strictly speaking, chaining the input space is

equivalent to running a completely new experiments, unless further assumptions are made to allow the transfer of correlations and information from other related dataset.

Kim E. Jelfs replied: You are absolutely correct. Changing the input space is equivalent to running a new campaign. The flexibility of Web-BO to facilitate this is facilitated by the database architecture. Storing "Experiments/Campaigns" and "Data" separately allows users to more readily adjust their parameter space, while taking advantage of the data that they have collected over the course of the optimization trajectory.

Ken Sakaushi queried: Does your BO system have a special strategy to proceed Bayesian optimization, such as simplifying a process for hyperparameter tuning?

Kim E. Jelfs replied: As Web-BO is in the initial development stage, we rely on the functionality and features that are offered by existing packages for Bayesian optimization in chemistry (namely, BayBE from Merck). As such, we do not offer special strategies to accommodate hyperparameter tuning; however, as these are implemented in the packages that Web-BO is built on, we will integrate the additional functionality accordingly.

Alán Aspuru-Guzik remarked: Helping experimentalists remove "unfeasible" regions of space (not synthesizable or failed experiments) is a useful idea. Another layerable addition to your web interface that would be cool is the use of a classifier followed by a Bayesian optimizer. The classifier is in charge of discovering the synthesizable regions. We have made progress towards this in our Anubis algorithm,[1] which is readily available to use as a plug-in with any BO algorithm of your choice.

1 R. Hickman, M. Aldeghi and A. Aspuru-Guzik, Anubis: Bayesian optimization with unknown feasibility constraints for scientific experimentation, *ChemRxiv*, 2023, preprint, DOI: **10.26434/chemrxiv-2023-s5qnw**.

Austin M. Mroz responded: Ensuring feasible experiments are suggested and removing "unfeasible" regions is essential to practical implementations of BO for experimental chemistry. This is a great suggestion, and one that we look forward to including our framework. Indeed, integrating algorithms of this nature is a development priority for us.

Ruiqi Wu communicated: How does Web-BO handle the tuning of hyperparameters for the Gaussian-process model, and what impact does this tuning have on the accuracy and efficiency of the optimisation process?

Kim E. Jelfs communicated in reply: This feature is not currently implemented in Web-BO. However, it is an important consideration and one that we are keen to integrate.

Ruiqi Wu communicated: Given the success of Web-BO in optimising organic reactions, what would you expect its performance to be for inorganic reactions?

Could Web-BO effectively determine the ideal reaction conditions for inorganic systems?

Kim E. Jelfs communicated in reply: Web-BO is an overarching GUI that allows one to access Bayesian optimisation without coding; it is not specific to any specific application area, so we would indeed expect it to perform as well for inorganic systems, and it is already set up to do this. We used organic reactions as the example in the paper (**https://doi.org/10.1039/d4fd00109e**) as there are open-source emulators available for these.

Barnabas A. Franklin communicated: Web-BO seems like a great tool for learning and experimenting with Bayesian optimisation. Do you have any plans to adapt this to other optimisation techniques?

Kim E. Jelfs communicated in reply: Absolutely, with the framework in place it is possible to add in additional optimisation options, and indeed additional Bayesian optimisers also.

Ian Fairlamb communicated: Do you think it will be possible to feed mechanistic information into the Bayesian optimisation tool; for example, how a specific Pd precatalyst is activated by a chemical trigger(s) under the given reaction conditions? We find that many cross-coupling reactions never get the chance to get going, especially when run using high-throughput experimentation (HTE) methods. This then leads potentially to false negatives, *i.e.*, a particular phosphine ligand is not active under the reaction conditions, when the reality is that it never had a chance to get going to assess the true ligand efficacy.

Kim E. Jelfs communicated in reply: Bayesian optimization (BO) algorithms featuring Gaussian processes (GPs) as the surrogate models may afford this feature; GPs are quite flexible and can be conditioned on a diverse array of prior data/information/beliefs about the system. Specific to the cross-coupling case, if reaction conditions are a part of the optimization campaign, these results would not necessarily be false negatives. Instead, the results indicate that the ligand is inactive under the conditions that it was run (including time). Where the power of BO comes into play here is the balance between exploitation and exploration; the algorithm formulation may be tuned (and constraints included) to encourage a more explorative search of the reaction conditions, which may result in improved performance.

Ferdinand Krammer communicated: You mentioned in response to another comment how universities are unwilling to pay for the infrastructure/hardware associated with the running and storing of the data needed for these optimisation approaches. I was wondering how you envisage it going forward, particularly with respect to experimental groups being unable to see the benefit behind creating their own hardware setups (especially if they have never had them before).

Kim E. Jelfs communicated in reply: Obviously storing data has associated costs. We have therefore made it possible (and likely the most common scenario)

for the user to download the software/app locally. This would not require the user to have specialist or complex hardware setups as Web-BO could be run on a standard desktop computer.

Matthew R. Ryder communicated: Web-BO seems like a great tool that promises to lower the barrier for non-expert users in Bayesian optimization. How do you envision incorporating real-world experimental feedback loops into this framework to validate and refine predictions iteratively, as complexity increases?

Kim E. Jelfs communicated in reply: This is an excellent point. Often, chemical challenges require more complex Bayesian optimization (BO) algorithm formulations, which are not currently supported by Web-BO. We are actively working on frameworks that allow researchers to more easily integrate these into their workflows. Within the context of closed-loop optimization (autonomous labs), we do not immediately envision Web-BO as playing an integral part in this strategy, as autonomous feedback loops require non-GUI-based solutions.

Christian Kuttner addressed Kim E. Jelfs: Democratizing the use of data-driven tools is an important topic. Concerning barriers to accessibility of data-driven tools in chemistry, how does a GUI-based BO platform help to overcome these technical barriers, particularly for experimentalists with limited ML expertise and computational skills?

Kim E. Jelfs answered: Providing the GUI means no coding or ML expertise is needed *a priori* by the user, so this directly targets this problem. We feel combining exposing users to the tool alongside some training on the background of Bayesian optimisation, how it works, and the importance of various choices (*e.g.* surrogate models, acquisition functions, *etc.*) is therefore the right approach.

Roohollah Hafizi opened the discussion of the paper by Arya Changiarath: Regarding the active learning process, it appears that the acquisition function (eqn (1) in your paper [**https://doi.org/10.1039/d4fd00099d**]) is used for this purpose. In this equation, σ represents the typical uncertainty measure commonly employed in active learning. Could you clarify the role of μ as the exploitation term in this context?

Arya Changiarath responded: By including μ, the algorithm is encouraged to select points that are expected to perform well based on current knowledge. This balances with σ, which promotes the exploration of uncertain areas.

Kevion K. Darmawan commented: When reading your paper (**https://doi.org/10.1039/d4fd00099d**), it is mentioned that one of the limitations of this work may be excessive hydrophobicity. Could you please suggest why this might be the case? How does the current CALVADOS force field compare to the MARTINI force field? Additionally, are you considering using all-atomistic systems for future machine learning models, given their enhanced structural resolution compared to coarse-grained models?

Arya Changiarath responded: In our model, we maximize the second virial coefficient ($|B_{22}|$) to predict strongly interacting peptides. However, in phase-separated systems, too much hydrophobicity can cause overly strong self-interactions. Optimising interaction strength requires balancing other factors like transport properties. A multi-parameter approach that balances hydrophobic and non-hydrophobic residues could create better-behaved protein sequences. While we have not used multi-parameter optimisation yet (as in An *et al.*),[1] it would likely be useful for designing new materials based on protein phase separation. CALVADOS is a residue-level coarse-grained implicit solvent model, which is optimised for modelling protein phase separation and condensate formation for intrinsically disordered proteins, and it excels in the detailed simulation of phase-separated condensates. MARTINI3 is also coarse-grained, but it groups atoms into larger beads and includes explicit solvents. The MARTINI3 force field has been revised to accurately model intrinsically disordered proteins (IDPs) but its ability to describe sequence-dependent effects on protein phase separation needs to be more thoroughly explored, and once its strengths and weaknesses are better established it will be an excellent model to study sequence–property relationships. CALVADOS is shown to provide accurate predictions for IDPs in terms of conformational properties and phase separation behaviour. Extending the insights gained from coarse-grained models like CALVADOS2 to all-atom models could lead to enhanced structural resolution. While coarse-grained models reduce computational costs and can still capture phase behaviour, all-atom models provide a detailed representation of molecular interactions. This could provide better input for the machine learning model by accounting for fine-scale details when greater accuracy in capturing protein dynamics and interaction specificity is required.

1 Y. An, M. A. Webb and W. M. Jacobs, Active learning of the thermodynamics-dynamics trade-off in protein condensates, *Sci. Adv.*, 2024, **10**, eadj2448, DOI: **10.1126/sciadv.adj2448**.

Nadine Schneider asked: Is it possible to validate *via* experimental data? Are there experimental data sets available where you can compare how good the model is? How reliable are the simulations compared to experiments?

Arya Changiarath responded: In this work, there is no direct comparison to experimental data for validating the computational model and simulations. However, the active learning model developed here has the potential to be validated against experimental data in future studies. This validation will help establish the reliability and real-world applicability of the predictions made by the model. The coarse-grained simulations employed in this study capture important experimental observations noted in previous research.[1]

1 A. Changiarath, D. Flores-Solis, J. J. Michels, R. Herrera Rodriguez, S. M. Hanson, F. Schmid, M. Zweckstetter, J. Padeken and L. S. Stelzl, Promoter and Gene-Body RNA-Polymerase II co-exist in partial demixed condensates, *bioRxiv*, 2024, preprint, DOI: **10.1101/2024.03.16.585180**.

Matthew R. Ryder communicated: Your use of active learning to explore large sequence spaces is impressive. Do you foresee limitations in applying this to more

complex systems that involve disorder, or even systems that feature phase transitions? How might this be addressed in future iterations?

Arya Changiarath communicated in reply: Studying the phase behaviour of intrinsically disordered proteins is challenging due to their high degree of conformational dynamics and complex interactions. This complexity makes it difficult for active learning models to capture the full range of behaviours and interactions across a wide sequence space. In our research (**https://doi.org/10.1039/d4fd00099d**), we examined the phase transition of two different peptides demixing to form multiphasic structures, using labels that capture peptide interaction strength and indirectly studying phase separation. To extend this work, we could optimize the saturation concentration (c_{sat}) or interfacial tensions of protein condensates, as demonstrated by P. Y. Chew *et al.*,[1] providing more direct insights into protein phase separation. Combining active learning with other techniques such as transfer learning and incorporating multiple features in the optimization process could help capture more details from the sequence space to develop a more robust and comprehensive model.

1 P. Y. Chew, J. A. Joseph, R. Collepardo-Guevara and A. Reinhardt, Thermodynamic origins of two-component multiphase condensates of proteins, *Chem. Sci.*, 2023, **14**, 1820–1836, DOI: **10.1039/d2sc05873a**.

Katharina Ueltzen opened the discussion of the paper by Yuchen Lou and Alex M. Ganose: The anisotropy ratio is not necessarily computed from the same eigenvalues in reference and prediction. Did you also evaluate whether your model predicts the dielectric tensor qualitatively right, *i.e.*, whether it predicts correctly in which crystallographic direction the dielectric response is the largest/the smallest?

Yuchen Lou replied: This is an excellent point and not something we explored in our work. For many crystals, the directions of the eigenvectors will be constrained by the crystal symmetry (*e.g.*, tetragonal, orthorhombic, hexagonal, *etc.*). However, even within these systems, the ordering of the eigenvalues could differ. We are currently working on an improved metric for capturing anisotropy that can account for this point.

Xinwei Wang remarked: Predicting the dielectric tensor is highly valuable, especially when considering practical applications. For example, it can serve as a metric to screen promising solar cell absorbers. On the one hand, as you showed, we can predict optical absorption based on dielectric tensors. On the other hand, high dielectric constants suggest a strong capacity to screen charge defects and reduce recombination losses. For the full static dielectric tensor, the ionic contribution typically dominates. Could you comment on the challenges involved in predicting this part? I am aware of some studies on predicting isotropic ionic dielectric constants using classical machine learning approaches,[1,2] but what do you think are the potential limitations when attempting to extend those models to predict the anisotropic ionic tensors?

1 A. Takahashi, Y. Kumagai, J. Miyamoto, Y. Mochizuki and F. Oba, Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations, *Phys. Rev. Mater.*, 2020, **4**, 103801, DOI: **10.1103/PhysRevMaterials.4.103801**.
2 Y. Hu, M. Wu, M. Yuan, Y. Wen, P. Ren, S. Ye, F. Liu, B. Zhou, H. Fang, R. Wang, Z. Ji and R. Huang, Accurate prediction of dielectric properties and bandgaps in materials with a machine learning approach, *Appl. Phys. Lett.*, 2024, **125**, 152905, DOI: **10.1063/5.0223890**.

Yuchen Lou responded: We have tried to use the architecture to predict the ionic dielectric tensor; however, the model did not perform very well. Ionic contributions depend not only on the atomic coordinates but also on the vibrations of the atoms, which are not trivial to obtain.

Sophie Bennett asked: Do you think this method will be generally applicable to higher rank tensors? Do you envisage any challenges?

Yuchen Lou replied: Yes! Ref. 1 used a similar architecture as AnisoNet to predict elastic tensors for crystalline materials, and the irreducible representations of elastic tensors include terms up to $L = 4$.

1 M.Wen, M. K. Horton, J. M. Munro, P. Huck and K. A. Persson, An equivariant graph neural network for the elasticity tensors of all seven crystal systems, *Digital Discovery*, 2024, **3**, 869–882, DOI: **10.1039/d3dd00233k**.

Gillian James commented: We can see that the equivariant graph neural network (GNN) data representation dramatically improves anisotropic dielectric tensor prediction. Is this exclusively a result of the embedded symmetry constraints of equivariant GNNs, or are there additional emergent properties of the GNN representation that improves the tensor-prediction? In other words, if there was some way to enforce the symmetry constraint on a scalar data representation of the dielectric tensor, would you obtain/reproduce the same prediction accuracy? Why or why not?

Yuchen Lou replied: GNNs have emerged as a natural representation for modelling crystals and molecules. While other approaches have been trialled, a key challenge is that they are not permutation invariant. Undoubtedly this helps to improve the tensor prediction accuracy; however, we have not compared directly against other structural representations.

Gillian James said: You point out that your model predicts anisotropic dielectric tensors for a lot of geometrically anisotropic materials. For example, many of the materials with high predictions of dielectric anisotropy are 1D or 2D materials, and you point out that even most of the 3D materials are 'pseudo-2D'. You also pointed out that many of the 2D materials with predicted anisotropic dielectric tensors were transition-metal dichalcogenides, but we already know that these chemistries frequently form 2D materials (*i.e.*, V_2O_5, TaS_2, *etc.*). Is your model only capturing geometric contributions to dielectric anisotropy or is it actually capturing chemical contributions to highly anisotropic dielectric tensors? Can you more-directly analyze chemical trends in dielectric anisotropy or are you able to explore this with your dataset?

Yuchen Lou responded: We believe AnisoNet is capturing both geometric and chemical contributions. For example, our dataset includes 2H-structured WS_2 (mp-224), WSe_2 (mp-1821), MoS_2 (mp-1018809), $MoSe_2$ (mp-1634). While all these materials have the same structure, the anisotropy ratio varies from X–X, and the polycrystalline dielectric constant varies from Y–Y, indicating that chemical features are playing a role.

Vishank Kumar asked: Can you explain how you explored the new potential chemical space? How did you generate new crystal structures? Or are these the materials in the Materials Project (MP) database for which the properties were not calculated?

Yuchen Lou replied: That is correct, the new materials were from the MP database but without a DFPT-calculated dielectric tensor.

Alán Aspuru-Guzik remarked: Your work would be more flexible if you turned your model into a *generative* model rather than a discriminative model that is useful only for screening. The model should be well conditioned. Have you and your collaborators considered moving in this direction? The conditioning, if done with stability in mind, would allow you to use your model with stability in mind from simple considerations such as the above-the-convex-hull energy or other more sophisticated synthesizability models. We have worked on generative models for solids early on and it was easy to condition on the energy;[1] you may want to look at our work in this area and that of others.

1 J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik and Y. Jung, Inverse Design of Solid-State Materials via a Continuous Representation, *Matter*, 2019, **1**, 1370–1384, DOI: **10.1016/j.matt.2019.08.017**.

Yuchen Lou answered: That is a good idea. We appreciate the suggestion and agree that it shouldn't be too hard to include. We can also apply classifier-free guidance on it to amplify the anisotropy of the generations.

Volker L. Deringer said: Following up on the previous question, I am also curious how your approach could become part of wider-ranging materials discovery initiatives. Could it be interfaced, for example, to crystal-structure prediction methods such as AIRSS (**https://doi.org/10.1039/d4fd00134f**) and FUSE (**https://doi.org/10.1039/d4fd00094c**) that were discussed in the previous session on 'Discovering chemical structure' (**https://doi.org/10.1039/d4fd90061h**)?

Alex M. Ganose answered: We can see these methods being used as part of a downstreaming screening pipeline that gets applied to the lowest-energy structures that emerge from AIRSS and FUSE studies. One issue with this approach is that you cannot push the predictions towards materials with more anisotropic properties. This is potentially a benefit of generative ML approaches where conditional generation can be baked into the model.

Yuchen Lou responded: This is an interesting idea and we see no technical barriers to achieving it. Inference with AnisoNet is fast – on the order of 1000 structures per second – so it should be possible to screen all generated structures for their dielectric properties.

Christopher M. Collins added: A potentially powerful option would be to use promising structures which emerge from ML models to then seed crystal structure prediction runs with FUSE and/or AIRSS. In practice, this can work exactly as set out in my paper (**https://doi.org/10.1039/d4fd00094c**), but additionally using structures directly out of your models in addition to those from the ML generator in the paper.

What this will do is take a structure that you think may be good for a given property and assess how stable it is within a composition, increasing the likelihood of discovering a material that has a target property.

Recently, we have done something in this area, where we were interested in compositions that are far away from known compounds in chemical space, which we enumerated with an automated reasoning model, and then used FUSE and ML property prediction to assess a subset of them for Li-ion conductivity as well as their stability.[1]

1 J. Clymo, C. M. Collins, K. Atkinson, M. S. Dyer, M. W Gaultois, V. Gusev, M. J. Rosseinsky and S. Schewe, Exploration of Chemical Space through Automated Reasoning, *Angew. Chem. Int. Ed*, 2024, e202417657, DOI: **10.1002/anie.202417657**.

Aron Walsh said: From a materials chemistry perspective, the anisotropy in the dielectric response may be linked to differences in chemical bonding along different crystallographic directions. Do you think taking a pre-trained force field model that has learned asymmetry in energies and forces would be effective for transfer learning or building models for multi-property prediction?

Yuchen Lou replied: This is an interesting suggestion and, in fact, was recently adopted in a related work to ours.[1]

1 Z. Mao, W. Li and J. Tan, Dielectric Tensor Prediction for Inorganic Materials Using Latent Information from Preferred Potential, *arXiv*, 2024, preprint, arXiv:2405.09052 [cond-mat.mtrl-sci], DOI: **10.48550/arXiv.2405.09052**.

Ksenia R. Briling communicated: Have you tried to learn and predict the polycrystalline dielectric constant and anisotropy ratio, or some other interesting experimental measurement, directly (output $2 \times 0e$) instead of computing them from predicted tensors? Do you think the tensor approach is better? (I would expect so, especially for extrapolation/discovery tasks.)

Yuchen Lou communicated in reply: While this would theoretically be possible, it will unnecessarily constrain the use of the model. For example, one would not have access to the specific directions in which the material is anisotropic meaning materials insights are limited.

Zsuzsanna Koczor-Benda communicated: Fig. 5a of your paper (**https://doi.org/10.1039/d4fd00096j**) shows that AnisoNet is less accurate for materials with a high

anisotropy ratio, due to the imbalanced training set. Have you tried to retrain the model on the 137 newly discovered structures with high anisotropy ratios to improve its accuracy and see if there are additional promising materials that were missed in the first screening round due to having an anisotropy ratio that was severely underestimated by the model?

Yuchen Lou communicated in reply: No we did not, but we appreciate the suggestion and believe it would be beneficial for improving the accuracy of our model.

Christian Kuttner communicated: Anisotropic dielectric materials are suggested as a platform for advanced applications. Out of curiosity, what specific aspects of their anisotropic properties make them suitable for these applications, and how might the discovery of highly anisotropic materials impact emerging technologies like metamaterials and thermoelectrics?

Alex M. Ganose communicated in reply: The impact of anisotropy varies for each application but to give some general ideas: For thermoelectrics, it would be highly beneficial if you can design a material with orthogonal directions for heat and charge transport. This should be useful in increasing the figure of merit. For birefringence, anisotropy of the optical response enables their application in displays and medical diagnostics. There are also emerging applications for anisotropic materials as dark matter detectors. If the interaction with dark matter is direction dependent, then this can be used as a signature based on the orientation of the detector relative to the proposed dark matter wind.

Graeme M. Day communicated: Can you comment on the increase in the amount of training data required for training a model to predict tensorial properties compared to scalar properties, such as the dielectric tensor *vs.* the polycrystalline dielectric constant using the same model architecture? Did you look at the learning rate, error *vs.* training set size, for both properties and how this compares?

Yuchen Lou communicated in reply: Thank you for the interesting question. We believe the results in our paper (**https://doi.org/10.1039/d4fd00096j**) go some way toward answering this question. We demonstrate that an equivariant model trained to predict the full dielectric tensor achieves comparable accuracy as a scalar model on the polycrystalline dielectric constant. Accordingly, the amount of training data needed for predicting the full dielectric tensor does not appear to be substantially more than for the scalar value. A further test could be to explore the learning curve of scalar and anisotropic properties with training dataset size. In the very first stage of this project, we trained the model on a dielectric constant dataset (*i.e.*, output shape $= 1 \times 0e$), and later on the dielectric tensor dataset of the same size (output shape $= 1 \times 0e + 1 \times 2e$) and the performances were not too different from each other. So we can tell that it wasn't because our data labels had 9 numbers *vs.* just a constant. (You gave me a test dataset of 4700 materials with constant labels right and that model had like 0.41 MAE, and I can say yes we are still well in the "more data the better" range).

Matthew R. Ryder communicated: You've shown promising results in discovering anisotropic dielectric crystals using equivariant graph neural networks. How do you plan to incorporate thermodynamic considerations, such as disorder or environmental effects, into the model to enhance its prediction accuracy in real-world applications?

Yuchen Lou communicated in reply: It is not trivial to apply these additional effects. For disorder, one could consider using the average of two elemental embeddings; however, to our knowledge this has never been trialled. For environmental effects, the recently developed unified differentiable approach to dielectric tensor prediction may be useful under applied electric fields.[1]

1 S. Falletta, A. Cepellotti, A. Johansson, C. W. Tan, A. Musaelian, C. J. Owen, and Boris Kozinsky, Unified Differentiable Learning of Electric Response, *arXiv*, 2024, preprint, arXiv:2403.17207 [cond-mat.mtrl-sci], DOI: **10.48550/arXiv.2403.17207**.

Bastian Bjerkem Skjelstad opened the discussion of the paper by Heather J. Kulik: I found your analysis of the metal-local environment very interesting, since it reveals structural motifs that are commonly used for certain classes of transition-metal complexes. Do you have any thoughts about how this information may be leveraged to design new transition-metal complexes; for example, more efficient catalysts than currently existing ones, in hitherto less explored areas of chemical space?

Heather J. Kulik responded: That's an interesting direction we haven't tried yet. One thought would be to preserve the metal-local features we have identified but then use a generative model or genetic algorithm to vary more peripheral features and enhance diversity, *e.g.*, in a large multi-objective optimization. I think we'd need to look at something broader than just a single catalytic reaction unless we were really stringent in our definition of the motifs included (*i.e.*, not just the whole tmCAT set). But we could probably do this most easily for spin crossover design, which was one of the smallest sets.

Bastian Bjerkem Skjelstad asked: Out of the 86 665 transition-metal complexes in the tmQM dataset, a total of 29 995 complexes were assigned to either the tmCAT, tmPHOTO, tmBIO or tmSCO dataset. Do you have any insights into what the remaining unassigned structures might be? Could they not be classified, or did they not fit any of the four curated datasets?

Heather J. Kulik replied: We found a large number of other clusters when we did the unsupervised learning – around 20. So these other clusters were just either more ambiguous, belonged to a smaller grouping, or were harder to assign. For example, one category was molecules that just referred to some degree of structural characterization, such that the paper itself might have been a crystallography paper – that would correspond to a less "interesting" molecule. Still others could have just been hard to assign based on the text. Further refinements could likely be made to our approach as language models evolve to hopefully assign more of these structures.

Matthijs Hakkennes remarked: For the tmBIO dataset, did you also extract information about biological targets, such as pIC_{50} or p*K* values?

Heather J. Kulik answered: We did not extract additional property values, in part because we were working directly with titles and abstracts rather than the full texts of the papers. Thus, we were limited in the types of quantities we could extract.

Adarsh V. Kalikadien commented: It was mentioned that it is hard to distinguish catalysis and non-catalysis complexes between the tmCAT and tmQM databases. It is discussed that these databases mainly contain precatalyst complexes, of which the descriptors are likely not related to the active catalyst species. Are you aware of any studies that have investigated this relation between precatalysts and active catalyst species based on a similar descriptor set or are you planning to do it?

Heather J. Kulik replied: That's an interesting question – I am not aware of any study that actively identifies in a systematic way the difference between precatalysts and catalysts. Motivated by one of our sponsored projects, we are currently trying to use knowledge from the CSD to predict how an arbitrary ligand coordinates to a metal with a message-passing neural network. We are then trying to explore combining possible combinations of precatalysts and replacement ligands to propose what active structures are. While in the general case, it may be easy to identify obvious solvents that could be removed from a precatalyst to make it active, there are many cases where the most reactive species that is forming *in situ* is unknown. Thus, I think a chief limitation for anything based on the CSD, including generative or other data-driven models, is that we may not be learning the right structures if we are interested in catalysts.

Adarsh V. Kalikadien asked: In experiments, would you expect the complexation process of a metal–ligand complex to be more dependent on the reaction conditions or the nature of the ligand structure? What was your expectation about the ability to differentiate between catalysis and non-catalysis structures in tmCAT and tmQM using your descriptor set?

Heather J. Kulik answered: Recently in closely related work we examined hemilabile ligands which can preferentially bind a metal in different poses.[1] Experimentally, these variations have been captured in the CSD in crystal structures, but I can imagine that there may be even more variation that occurs in solution than we know about, including transiently forming multi-metallics. Certainly we might expect that there is some variation in how a ligand complexes a metal, but from what we've seen, around 90% of ligands only bind in one pose in crystal structures and so I would naively expect that a lot can be determined just based on mining crystal structures. Nevertheless, more to your second question, we thought there might be a difference in the steric bulk between catalysis and non-catalysis complexes. We did not really expect any substantive difference in electronic properties such as the HOMO–LUMO gap or metal partial charge. What surprised us was that there was no difference in the

buried volumes, despite having a clear difference in the favored symmetries of the complexes.

1 I. Kevlishvili, C. Duan and H. J. Kulik, Classification of Hemilabile Ligands Using Machine Learning, *J. Phys. Chem. Lett.*, 2023, **14**, 11100–11109, DOI: 10.1021/acs.jpclett.3c02828.

Filip T. Szczypiński remarked: Transition-metal complexes categorised as "non-catalytically active" might not only originate from catalyst precursors. It is very common to deposit organometallic and supramolecular crystal data purely for structural investigations and because of the aesthetic appeal of the structures. Your datasets might help us identify structures that could exhibit promising catalytic activity – *e.g.*, by their proximity to the clusters of known active compounds – but have never actually been tested with such functionality in mind. It is potentially an incredibly useful collection of data labels for dataset exploration.

I was also wondering what the rationale was behind limiting the overall charge to +1. Such a criterion can easily exclude many interesting polymetallic species and structures where counterions have been removed during data curation.

Heather J. Kulik replied: Thank you for your comment! Regarding the charge being limited to +1/−1 – that is the assumption in the DFT data that we used (but did not curate ourselves). This could include structures that have higher or lower charge but have been assigned these more neutral charges in the DFT calculations. We agree that for the DFT data to be useful, it should be curated at the charge state more reflective of the experimental complex. We are currently pursuing new data curation along those lines.

Nicholas David asked: Why are the earth-abundant transition metals less prevalent in your catalysis dataset? Does it have to do with the time window (likely 2000 to now) imposed by the limitations of text mining? Is this related at all to the synthesizability of these more earth-abundant transition-metal catalysts? For example, there exists an overabundance of publications related to graphene, partly due to the scotch tape method for synthesizing graphene flakes (along with graphene's intriguing electronic properties), which makes researching graphene highly accessible.

Heather J. Kulik answered: I wouldn't say that earth-abundant transition metals are necessarily less abundant, they are just not over-represented in the catalysis subset (or may be under-represented). I suspect that, rather than synthesizability, this could be related to their reactivity – they may be harder to isolate in a stable crystal. For example, Fe(IV)=O is a well-known oxidative moiety, but one will never see Fe(IV)=O in a crystal structure because it is so fleeting (the resting-state Fe(II) may be present in the CSD though).

Christian Kuttner asked: You mentioned you had to opt for plan B. Could you briefly tell us something about plan A and what went wrong?

Heather J. Kulik replied: Our plan A was to explore the role of representation/model choice in learning transition-metal complex properties in order to

recommend a "best in class" representation for transition-metal complexes. What we found was that current datasets were sufficiently noisy that we could not make any straightforward recommendation because model errors are still much higher than those for organic datasets.

Rob Evans communicated: On the topic of long-term trends or other patterns observed in your analyses of the transition-metal complex literature, I was somewhat surprised to hear that no trends had been observed. Is this an issue of how the data has been visualised and illustrated? In any case, is the absence of any patterns or trends interesting on its own, given the implications for future investigations of less well-studied space?

Heather J. Kulik communicated in reply: We had discussed looking at trends over time in the data, but it was really hard for us to identify a clear quantitative trend that wouldn't be "cherry picking". Definitely the discussion from this meeting has motivated us to go back and look at whether there are more obvious trends we could explore. I think the most interesting thing to do would be to look at the rise and frequency of specific substructures – *i.e.*, to identify when a new substructure comes in that spawns many follow-up studies. That almost certainly would show some time dependence.

Rob Evans communicated: All of the information presented took the form of statements – what complex structure, what element, *etc.* Is it possible to use language models to curate the datasets according to numerical or quantitative data?

Heather J. Kulik communicated in reply: One shortcut we took in our literature curation approach was to only use the title and abstract, which are in the public domain and easily scraped. Given the advent of large language models, it has become increasingly challenging to get publishers to give access to full text. Nevertheless, we have previously used mechanical text parsing and language processing to extract properties (in this case, for MOFs). One key challenge is named entity recognition – *i.e.*, identifying which structure in a paper is associated with which property. If both of those potential impediments are overcome, it should be possible to associate properties with structures using language models as well.

Christian Kuttner communicated: The creation of distinct datasets for catalysis, photophysics, biology, and magnetism enables more focused computational screening. How could these refined datasets accelerate ML-based material discovery, and what challenges remain in applying these methods to more complex or multi-functional materials?

Heather J. Kulik communicated in reply: That's an interesting question. I think the first thing to do would be to curate more accurate properties separately on each of these sets. However another interesting thing to do would be to try to augment these datasets to the scale that they could be suitable for generative models (*i.e.*, for novel structures). Another idea that came up in a different question is to specifically exploit common metal-local motifs, again for

generating structures. I think the challenge for applying the same type of approach say to polymers or metal–organic frameworks is that we really have a very limited understanding of the concentration of defects, despite the fact that defects can influence properties strongly. As a result, when we would map out the relationship between literature-mined properties and the MOFs or polymers, we would have to implicitly account for this disorder without explicit knowledge of it (*e.g.*, for DFT-based modeling). I am not sure how to address that because I am not sure that people are always able to be quantitative about how present defects are in their materials.

Eltjo Mante communicated: Thank you for the paper, it was an incredibly interesting read. You mention that complexes which come from motifs identified to reside in the tmCAT, tmPHOTO, and tmBIO sets are commonly used as triplet sensitizers. Did this finding direct you to test other commonly used triplet sensitizers, as this would allow you to increase the size of three sets, with only one structure?

Heather J. Kulik communicated in reply: Our analysis does indeed suggest some classes of compounds have multiple types of applications. Pursuing triplet sensitizers and specifically curating a set of them based on language modeling would be one way to increase all the sets. It's an interesting suggestion but not one we have tried yet.

Eltjo Mante communicated: When you say "Analysis of common application-specific structural motifs reveals that, for a given motif, there are complexes within the tmQM dataset that contain the motif but the associated manuscripts do not indicate the complex has been assessed for that specific application." in your paper (**https://doi.org/10.1039/d4fd00087k**), have you tested some of the unassessed complexes to see whether they offer better performance for a specific application than what we currently have?

Heather J. Kulik communicated in reply: We have not yet tested these "similar" complexes to see if they have as good or better properties in comparison to the ones we could confidently label for a specific application. That's a great suggestion and would be something we would like to try out next.

Eltjo Mante communicated: In the Conclusions of your paper (**https://doi.org/10.1039/d4fd00087k**) you say "The curated tmCAT, tmPHOTO, tmBIO, and tmSCO datasets are expected to enable more focused high-throughput computational screening and development of predictive machine learning models while still allowing for exploration across diverse chemical spaces." Have you done any further screening on these datasets now, and if so, which ones have offered the most promising results?

Heather J. Kulik communicated in reply: We have not yet carried out calculations on these subsets. We think the key next step is to properly assign the oxidation state of the metal and overall charge of the complexes, as well as to probe putative spin states for mid-row transition metals. We are currently

working on refining our charge/spin assignment approach and are in the early stages of property curation.

Shubham Vishnoi opened the discussion of the paper by Daniel Crusius: In your opinion, how does incorporating noise into synthetic datasets, such as those generated through physics-based simulations, enhance a model's ability to predict experimental outcomes? Can the addition of noise improve the accuracy of predictions and help the model perform better when applied to real-world experimental data?

Daniel Crusius responded: Noise added to synthetic datasets will not necessarily enhance the model's ability to predict outcomes. However, adding noise to synthetic datasets could allow us to study how robust different models are to varying levels of noise, which could be an important learning process when, *e.g.*, pre-training on synthetic data and later fine-tuning with (noisy) experimental data.

Alán Aspuru-Guzik remarked: Noise distributions, for many reasons, can be non-Gaussian; for example, the general case of "noise in the noise". There is also the case of "uncertainty in the noise distribution". Are you planning to handle these cases in your approach?

Daniel Crusius replied: As a first simple case, we only considered Gaussian-distributed noise. However, different noise distributions can easily be added into our open-source Python package NoiseEstimator (**https://github.com/D-Cru/NoiseEstimator**), which we plan to expand in the future.

Alán Aspuru-Guzik asked: Have you considered making your structure a readily available "plugin" for other models in a well-documented, pip-installable fashion? This could help popularize your tool.

Daniel Crusius replied: The package is open-source and available at **https://github.com/D-Cru/NoiseEstimator**. It already is pip-installable and can be readily integrated into other models. For ease of use, the tool can also directly be used at **https://noiseestimator.bioch.ox.ac.uk**.

Tahereh Nematiaram commented: Your work effectively highlights the impact of noise on model performance and provides useful tools for setting realistic expectations. However, considering the variability in data quality across different sources, how can multi-source datasets be combined in a way that enhances model generalization without introducing bias or overfitting, particularly in fields where data is often limited or highly heterogeneous?

Daniel Crusius answered: Combining multi-source datasets will be a trade-off between increased experimental noise and a bigger applicability domain; the latter can indeed enhance a model's ability to generalise.

Kevion K. Darmawan queried: Compared to available programs, including Schrödinger's QikProp, how do the machine learning models predict drug

pharmacokinetic properties? Additionally, for future work, are you considering developing machine learning models for predicting blood–brain barrier permeability?

Daniel Crusius replied: Our study primarily focuses on how experimental noise impacts the predictive power of machine learning models, rather than developing or comparing specific machine learning techniques. Exploring how noise affects predictions from models like QikProp or developing specialised models for properties such as blood–brain barrier permeability would be interesting future directions, though they were not the focus of our current work.

Brett M. Savoie commented: One takeaway from this work is that if the model outperforms the experimental uncertainty of the data (aleatoric/irreducible error) then it is clearly learning noise or there is data leakage. This is useful as a diagnostic, but it doesn't rule out cases where the performance is comparable or even worse than the irreducible error but still learning noise. Are there any diagnostics or evaluation practices that would work in these cases?

Daniel Crusius answered: This is a very good point. Thorough evaluation practices can help diagnose if a model is learning noise or data leakage occurs, but depend on the specific application. One example when working with molecular datasets could be to rely on different kinds of data splits, such as comparing a random split to a scaffold-based split to capture the extent of data leakage/learning to noise.

Annabel Eardley-Brunt commented: This work considers noise in the labels, or the y data; have you thought about how it would handle errors in the input variables as well (the x data if considering a traditional-use x to predict a y machine learning problem)? This applies to instances where machine learning is not being used on structures as inputs, but rather experimentally or computationally derived values, such as physical properties, spectra, or energies.

Daniel Crusius answered: Errors in the x data is not something we have considered as part of this work, but could be especially interesting in the context of inorganic materials.

Chloe Wilson asked: Rather than trying to remove noise from datasets, would it be more useful to train an ML model which can identify which training data points are noisy and learn to ignore (*i.e.*, not fit to) these data points?

Daniel Crusius responded: This could also be a valid approach, however this would require detailed knowledge of which training data points to trust more, or repeat measurements. Without knowledge of which points are noisy and which ones are reliable, it would be an outlier detector, but depending on the specific application, outliers (such as activity cliffs) are desired and should not be ignored.

Christian Kuttner addressed Daniel Crusius and Sanggyu Chong: Your study (**https://doi.org/10.1039/d4fd00091a**) highlights that experimental errors in chemical datasets are often overlooked when assessing the performance of ML models, leading to over-optimistic interpretations of predictive power. How can

we encourage the adoption of more realistic metrics for model success? Perhaps Sangyu Chong would also like to comment on this.

Daniel Crusius communicated in reply: You make a very good point. Realistic metrics depend on the specific application area. One approach could be to define meaningful and standardised benchmarks (such as pre-defined data splits, evaluation metrics, *etc.*) that have real-world applicability and relevance. Additionally, statistical tests to evaluate if a model that scores better in a specific metric outperforms the model in a statistically significant way could aid model comparison.

Sanggyu Chong communicated in reply: From the prediction rigidity's point of view, noise or errors associated with the reference data, *i.e.*, the aleatoric error,[1] is not something that is explicitly considered in our formulations. It is, however, worthwhile to note the possibility of performing the "noise-aware" training of machine learning models, where the data points are weighted differently according to the level of their aleatoric error within the training set. I believe this can be one effective way of addressing the issue of intrinsic experimental errors, and when this approach is adopted, aleatoric errors can be indirectly addressed in the prediction rigidities by applying the said weights to the computation of the inverse Hessian for the training set.

1 E. Heid, C. J. McGill, F. H. Vermeire and W. H. Green, Characterizing Uncertainty in Machine Learning for Chemistry, *J. Chem. Inf. Model.*, 2023, **63**, 4012–4029, DOI: **10.1021/acs.jcim.3c00373**.

Sarbani Patra opened the discussion of the paper by Sanggyu Chong: Several methods are available for quantifying uncertainty in machine learning models of data. What are the specific advantages that prediction rigidity (PR) offers over its counterparts that might encourage one to select this method over other measures? How much additional effort in terms of computational time and expense would it entail over more simple measures? It would be immensely helpful if you could provide a brief illustration by comparing PR to one or more uncertainty quantification methods.

Sanggyu Chong responded: The advantages of adopting a prediction rigidity-based approach is largely two-fold. (1) It offers a very cheap way to obtain the uncertainties, without the need for any modifications to the model architecture or the loss function (which is commonly the case for any ensemble-based methods). All that is needed is the acquisition of the covariance matrix of the descriptors or last-layer features of the training set, and its inversion. For neural-network models, this is largely equivalent to one training epoch. Afterwards, a simple calibration of the metrics to a validation set is needed, then no additional computational cost is incurred thereafter. (2) It offers insights beyond uncertainties on the global predictions. In machine learning for chemistry, many models adopt a "locality" ansatz, where the model predictions are made locally for atoms or environments, then summed together to obtain the global prediction. Similarly, models can also adopt approaches where multiple prediction components (*e.g.* body-orders, multiple distance ranges) are combined together. While

these approaches have led to improved scalability, transferability, and interpretability, there have not been methods for assessing the model robustness at these levels where the raw, intermediate predictions are made. The PRs defined on these prediction levels have newly allowed for this, as demonstrated in our paper (**https://doi.org/10.1039/d4fd00101j**).

Chiheb Ben Mahmoud remarked: I understand that prediction rigidity (PR) is one way to look at the uncertainty of the ML models. However, it seems to be tightly related to the ML architecture and/or the features. How can we use this approach to build atomistic datasets efficiently, as in with the least amount of data? I think this question also links with yesterday's discussion (**https://doi.org/10.1039/d4fd90061h**) about how much data is needed to get a certain accuracy and how does that affect the predictive power of the models beyond their training data.

Sanggyu Chong replied: Thanks for raising this point. I agree that this is connected to some of the earlier discussions (yesterday's session [**https://doi.org/10.1039/d4fd90061h**], as well as some of the points discussed in this session for Prof. Kulik's paper [**https://doi.org/10.1039/d4fd00087k**]) on efficient dataset construction. The task of efficiently constructing a dataset to train a model of sufficient accuracy with the least amount of data is similar in spirit to the case study presented in Section 4 of our paper (**https://doi.org/10.1039/d4fd00101j**). Here, the PR would be calculated for a validation set (rather than a subset of interest) that spans the entire chemical space of interest for the model. Then, *from scratch* (*i.e.*, PR = 0 for all validation structures), a dataset can be composed by choosing, or generating, the structures that best increase the PR for all of the validation set structures. Structures would be added until a certain threshold value for the PR is met for all of the validation set structures. The key benefit of adopting this PR-based approach is that no model training or reference label acquisition is necessary; only the input parameters (*e.g.*, structural descriptors) are needed. For linear and kernel models, this is straightforward. For neural network-based models, one could try the same with the last-layer features of a randomly initialized model, motivated by the theory of neural tangent kernels.[1] I agree that structure selection would depend on the features or the ML architecture chosen. It would be very interesting to explore how much the dataset distribution differs from one architecture to another.

1 A. Jacot, F. Gabriel and C. Hongler, Neural Tangent Kernel: Convergence and Generalization in Neural Networks, in *Advances in Neural Information Processing Systems*, ed. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, 2018, vol. 31, pp. 8571–8580.

Niamh Hickey said: With molecular dynamics (MD) you can get many values and graphs; are you looking at extending this to other values we get from molecular dynamics? MD is computationally and time-intensive. Now that you have used this technique for looking into the pair correlation function and the Steinhardt order distribution, would you be interested in looking into other results that you can get from molecular dynamics?

Sanggyu Chong responded: In Section 6 of our paper (**https://doi.org/10.1039/d4fd00101j**), we extend our PR formalism to coarse-grained ML models and demonstrate the utility of PRs in assessing the trained coarse-grained ML models. From our experience, accurate evaluation of the models is only possible when MD is performed *with* the trained coarse-grained ML models and PR analysis is performed on the resulting trajectories. Our expectation is that looking into the relative PR simplifies a lot of the model validation process, especially in cases where it is unclear what analysis needs to be performed, or there are too many metrics to keep track of at once (*e.g.* many pair correlation functions). We agree that reaching sufficient convergence in these analyses can be time-intensive. However, as shown in Fig. 8 of our paper (**https://doi.org/10.1039/d4fd00101j**), deviation of the relative PR of the coarse-grained ML model trajectory from the reference value takes place very quickly, and hence obtaining a long MD trajectory may not be necessary for this PR-based assessment of the coarse-grained ML models.

Venkat Kapil commented: It can often be that many models are wrong 'in the same way'. I wonder if the prediction rigidities can quantify the bias.

Sanggyu Chong replied: This is a very interesting question. The prediction rigidities (PRs) are solely dependent on the distribution of the data points with respect to the training set. When it is used for uncertainty quantification, it is hence common practice to perform calibration of the metrics to ensure that the resulting error estimates are provided as accurately to the reference data as possible.[1] Bias could also be indirectly addressed during this calibration step. It is worth noting, however, that computing the prediction rigidities is analogous to performing a Laplace approximation, which makes a Gaussianity assumption. Extending the applicability of PR-based metrics to datasets with severe non-Gaussianity is one topic of our subsequent research efforts.

1 F. Bigi, S. Chong, M. Ceriotti and F. Grasselli, A prediction rigidity formalism for low-cost uncertainties in trained neural networks, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 045018, DOI: **10.1088/2632-2153/ad805f**.

Alán Aspuru-Guzik said: Hi Sanggyu, great talk on a very important topic! The machine learning community has a variety of well-established methods for uncertainty estimations. I cannot stop from seeing a solid similarity of your process to the well-employed Laplace approximation,[1] which is readily available using a "one-line-of-code" approach in Pytorch.[2] It would be great to explore, and benchmark against, these popular tools. My group is already using this approximation in our work to great success.

1 A. Kristiadi, M. Hein and P. Hennig, Learnable uncertainty under Laplace approximations, in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, ed. C. de Campos and M. H. Maathuis, Proceedings of Machine Learning Research, PMLR, 2021, vol. 161, pp. 344–353.
2 E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer and P. Hennig, Laplace Redux – Effortless Bayesian Deep Learning, *arXiv*, 2021, preprint, arXiv:2106.14806 [cs.LG], DOI: **10.48550/arXiv.2106.14806**.

Sanggyu Chong replied: Alán, many thanks for this important comment.

Indeed, our derivations of the prediction rigidities[1,2] that strive to quantify the robustness of model predictions result in expressions that can be considered largely equivalent to performing the well-established Laplace approximation. The added value offered by our metrics, which is the *only* aspect for which we would like to claim their novelty of, is the intuitive interpretation and analysis of the model predictions based on our metrics, which is especially useful when applied on the levels at which the ML models for chemistry makes their "raw" predictions (*e.g.* local/atomic or component-wise predictions that are not the actual prediction targets), as demonstrated in Sections 3 through 5 of our contribution (**https://doi.org/10.1039/d4fd00101j**). With regards to your second point, we are glad to hear that similar methods have been developed and adopted successfully in your group, and would be happy to explore and benchmark other relevant tools as you suggest. Based on our own experiences, we also agree with the easy-to-use nature of Laplace approximation-based uncertainty quantification methods, and hope that more researchers in our field will adopt such workflows and report their ML-based scientific findings with reliable error estimates.

1 S. Chong, F. Grasselli, C. B. Mahmoud, J. D. Morrow, V. L. Deringer and M. Ceriotti, Robustness of Local Predictions in Atomistic Machine Learning Models, *J. Chem. Theory Comput.*, 2023, **19**, 8020–8031, DOI: **10.1021/acs.jctc.3c00704**.
2 F. Bigi, S. Chong, M. Ceriotti and F. Grasselli, A prediction rigidity formalism for low-cost uncertainties in trained neural networks, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 045018, DOI: **10.1088/2632-2153/ad805f**.

Basita Das remarked: Noise depends on the experiment. It varies with experimental setups, and the person doing the experiment, not on the ML model.

Sanggyu Chong responded: Sure, and hence the ML community's efforts to better decouple the aleatoric and epistemic errors from one another. One should also be mindful of the fact that computational results can also contribute noise to the dataset, induced by factors such as poor convergence in the computations (*ab initio*) or the sampled thermodynamic ensembles (classical simulations).

Brett M. Savoie opened the discussion of the paper by Branko Ruscic: The logic behind the Active Thermochemical Tables (ATcT) is to revise provisional and independent experimental values to be self-consistent with thermodynamic relationships. Underneath this process, there is still a data curation step that requires judging amongst ostensibly very similar experiments and deciding the uncertainty of these values. For example, the paper (**https://doi.org/10.1039/d4fd00110a**) describes the evolution of many of the specific values and the role of trusted curators like Pedley, *etc.* Is there any way to do better than expert manual judgment at the data curation stage, or for the time being are there simply no alternatives?

Branko Ruscic replied: Let me start by depicting the traditional process, and then juxtapose it to the ATcT approach. The traditional sequential process of building a thermochemical database (A begets B, B begets, C, *etc.*) consists of a series of steps. Each step focuses on a particular target chemical species, during which the evaluator performs a literature search and identifies all thermochemically relevant measurements that connect the target species to those

determined in previous steps. The available determinations are critically evaluated and the 'best' measurement is selected and used to obtain the thermochemistry of the target species, treating the thermochemistry of all other species in the selected chemical reaction as *a priori* known. The evaluator then adds the thermochemistry of the target species to those considered known and moves to the next chemical species. The sequential procedure leads to several interesting problems. Probably the most detrimental is that the resulting tabulation contains hidden progenitor–progeny relationships, making future improvements very difficult or impossible. Namely, if later there appears a new determination that would lead to the change or improvement of the thermochemistry of one of the chemical species already listed in the tabulation, while perhaps nominally improving the thermochemistry of that species, it will introduce new inconsistencies across the tabulation, since the table contains other species whose thermochemistry was pegged to the old value of the just revised species, but these are not easily identifiable because of the hidden progenitor–progeny relationships. Other problems of the sequential approach are, for example, a growing cumulative error as the sequence evolves, incomplete use of the available information, *etc.*

The thermochemical network approach of ATcT completely resolves the problem of new or additional determinations: a new determination is simply added to the network and ATcT is rerun, propagating its consequences through all affected species. ATcT also gets rid of the cumulative error present in the sequential approach, and produces other numerous advantages, such as the covariance matrix, the ability to test hypotheses, identification of 'weak links' in the network, *etc.* The most relevant aspect related to your question is that ATcT, in fact, gets rid of the requirement to manually judge amongst ostensibly very similar experiments in order to select the 'best' available determination: all available determinations are entered into the thermochemical network and analyzed by the internal statistical analysis of ATcT and in the end contribute – with varying weights – to the final answer. One of the consequences is that the provenance of each ATcT value does not rely on a single measurement (as it does in sequential thermochemistry), but is distributed over a considerable number of determinations, which both increases the robustness and the final accuracy of ATcT results.[1] The detailed provenances of each ATcT value can be obtained by variance decomposition and are given for each species on the ATcT website.[2]

However, ATcT still requires a fair bit of curation in preparing the data that is used to construct the thermochemical network, such as literature searches to find thermochemically relevant papers, followed by a critical analysis of each paper. The latter involves figuring out what was the actually directly determined quantity (as opposed to the derived quantity, which may incorporate additional auxiliary thermochemistry taken by the authors from other sources), checking if the experiment or computation is sound enough and if the determination can be inserted into the thermochemical network as given by the authors, or if it requires a reinterpretation, a refit of the reported individual points, a re-computation of the measurement using more recent values for the constants, *etc.* The critical analysis also involves checking the nature of the authors' proposed uncertainty or its estimation if one is not given or is clearly overly optimistic in order to establish the initial ('prior' in Bayesian terminology) uncertainty entered into the thermochemical network. However, the curation does not involve direct comparison

to other papers that may have determined the same thermochemical quantity. For now, the initial curation of data that enter the thermochemical network is entirely manual work, at least until the stage at which ATcT software takes over and uses the prepared entries to construct, statistically analyze, and solve the thermochemical network. Perhaps in a more distant future we will be able to use ML to train AI models that will be able to perform literature searches, autonomously perform an expert critical analysis of each paper, extract the relevant determination, and assign a realistic prior uncertainty. Several times in the past, we have tried to discuss with our colleagues in the computer science area the possibility of building a system that would monitor our curation activities and eventually learn how to perform them autonomously, but each time it was concluded that the available tools are not yet sufficiently sophisticated for such an undertaking. However, we believe that enabling such capabilities is simply a matter of time.

1 B. Ruscic, D. Feller and K. A. Peterson, Active Thermochemical Tables: Dissociation Energies of Several Homonuclear First-Row Diatomics and Related Thermochemical Values, *Theor. Chem. Acc.*, 2014, **133**, 1415, DOI: **10.1007/s00214-013-1415-z**.
2 B. Ruscic and D. H. Bross, *Active Thermochemical Tables (ATcT) values based on ver. 1.202 of the Thermochemical Network*, Argonne National Laboratory, Lemont, Illinois, 2024, available at **https://ATcT.anl.gov**, DOI: **10.17038/CSE/2440256**.

Venkat Kapil remarked: I am really impressed by the approach. We predicted the free-energy differences of glycine polymorphs from machine learning potentials and thermodynamic integration[1] and were challenged by the small energy differences. In these cases, it becomes increasingly important to quantify error bars. I was surprised to see the large discrepancy in the literature for α-glycine formation energy. In these cases, I was curious if your approach predicts a final number (plus an error bar) by eliminating erroneous values or by reevaluating the error bars.

1 V. Kapil and E. A. Engel, A complete description of thermodynamic stabilities of molecular crystals, *Proc. Natl. Acad. Sci. U.S.A.*, 2022, **119**, e2111769119, DOI: **10.1073/pnas.2111769119**.

Branko Ruscic answered: Indeed, quantification of uncertainties is becoming increasingly important for a variety of reasons. One strong reason is that modelling of complex chemical environments is now reaching commendable levels of fidelity, and the researchers are approaching the point where it would be highly advantageous to be able to apportion the observed differences between the model predictions and targeted benchmarks to those that are entirely a consequence of residual uncertainties in the input parameters (such as, for example, thermochemistry and kinetic rate constants), those that are attributable to uncertainties in the benchmarks, and those that need to be attributed to residual imperfections in the model. Another example where uncertainties play an important role are small differences between thermochemical quantities, and your mention of the related challenge to reproduce them computationally is fully appreciated. Namely, in order to make small differences in estimated or computed values statistically relevant, the corresponding uncertainties need to be sufficiently small, and certainly smaller than the actual differences. As per the current ATcT results[1] and our paper (**https://doi.org/10.1039/d4fd00110a**), the 298.15 K enthalpies for the phase

transitions are: $\Delta_r H^\circ_{298}(\alpha\text{-glycine} \rightarrow \gamma\text{-glycine}) = -0.27 \pm 0.10$ kJ mol^{-1} and $\Delta_r H^\circ_{298}(\beta\text{-glycine} \rightarrow \gamma\text{-glycine}) = -0.59 \pm 0.13$ kJ mol^{-1}, resulting in the following free energies: $\Delta_r G^\circ_{298}(\alpha\text{-glycine} \rightarrow \gamma\text{-glycine}) = -0.17 \pm 0.10$ kJ mol^{-1} and $\Delta_r G^\circ_{298}(\beta\text{-glycine} \rightarrow \gamma\text{-glycine}) = -0.30 \pm 0.13$ kJ mol^{-1}. The listed ATcT values rely on the experimental calorimetric measurements of Perlovich, *et al.*[2] and Drebushchak, *et al.*[3] and have been determined as differences between calorimetric determinations of the solution enthalpies of each polymorph. We should note two important aspects here. First, the ATcT uncertainties are sufficiently small, making the reported ATcT enthalpies and free energies of the phase transition relevant. Second, of the three polymorphs, β-glycine is the least stable and thus $\Delta_r G^\circ_{298}(\beta\text{-glycine} \rightarrow \gamma\text{-glycine})$ is more negative than $\Delta_r G^\circ_{298}(\alpha\text{-glycine} \rightarrow \gamma\text{-glycine})$. While the computational approach from ref. 4 is clearly an important achievement, its predictions for glycine (Fig. 4 of ref. 4) seem to be a bit off, both in magnitude and in the relative order of the α and β polymorphs. I have no doubt that future developments of your approach will enhance the computational fidelity to the point of making both the computational uncertainty sufficiently small and α-glycine more stable than β-glycine.

As per your second comment, we were also rather surprised by the scatter in the calorimetric enthalpies of formation of α-glycine. The reason we have opted to use glycine as an illustration of the capability of ATcT to successfully arbitrate between experimental values using a thermochemical network is that in this particular case the manual critical analysis of literature calorimetric experiments lacks convincing arguments for identifying errant values, other than – perhaps – relying on the general track record of individual researchers. It should be mentioned here that, unlike the typical manual procedure of critical evaluation, ATcT does not outright eliminate values that appear erroneous. Rather, as mentioned in our paper (**https://doi.org/10.1039/d4fd00110a**) (and explained in more detail in the papers that originally introduced ATcT),[5,6] ATcT uses statistical analysis to identify and iteratively augment the uncertainty of any determination that seems inconsistent with the prevailing knowledge content of the underlying thermochemical network, effectively gradually lowering the influence of the errant determination.

1 B. Ruscic and D. H. Bross, *Active Thermochemical Tables (ATcT) values based on ver. 1.202 of the Thermochemical Network*, Argonne National Laboratory, Lemont, Illinois, 2024, available at **https://ATcT.anl.gov**, DOI: **10.17038/CSE/2440256**.
2 G. L. Perlovich, L. K. Hansen and A. Bauer-Brandl, The Polymorphism of Glycine: Thermochemical and Structural Aspects, *J. Therm. Anal. Calorim.*, 2001, **66**, 699–715, DOI: **10.1023/A:1013179702730**.
3 V. A. Drebushchak, Y. A. Kovalevskaya, I. E. Paukov and E. V. Boldyreva, Low-temperature heat capacity of α and γ polymorphs of glycine, *J. Therm. Anal. Calorim.*, 2003, **74**, 109–120, DOI: **10.1023/A:1026377703260**.
4 V. Kapil and E. A. Engel, A complete description of thermodynamic stabilities of molecular crystals, *Proc. Natl. Acad. Sci. U.S.A.*, 2022, **119**, e2111769119, DOI: **10.1073/pnas.2111769119**.
5 B. Ruscic, R. E. Pinzon, M. L. Morton, G. von Laszewski, S. J. Bittner, S. G. Nijsure, K. A. Amin, M. Minkoff and A. F. Wagner, Introduction to Active Thermochemical Tables: Several "Key" Enthalpies of Formation Revisited, *J. Phys. Chem. A*, 2004, **108**, 9979–9997, DOI: **10.1021/jp047912y**.
6 B. Ruscic, R. E. Pinzon, G. von Laszewski, D. Kodeboyina, A. Burcat, D. Leahy, D. Montoya and A. F. Wagner, Active Thermochemical Tables: Thermochemistry for the 21st Century, *J. Phys.: Conf. Ser.*, 2005, **16**, 561–570, DOI: **10.1088/1742-6596/16/1/078**.

Volker L. Deringer commented: In the field of machine-learned interatomic potentials, there is a major challenge in validating ML potential models, both in terms of numerical quality and of how they predict physical properties.[1] Your contribution showed a treasure trove of high-quality, carefully curated experimental thermochemistry data – could these allow you to build a benchmark set on which ML potentials could be tested?

1 J. D. Morrow, J. L. A. Gardner and V. L. Deringer, How to validate machine-learned interatomic potentials, *J. Chem. Phys.*, 2023, **158**, 121501, DOI: 10.1063/5.0139611.

Branko Ruscic answered: An excellent question! The ATcT thermochemistry was, in fact, used for the development and benchmarking of several state-of-the-art composite electronic methods, capable of sub-kJ mol^{-1} accuracies for smaller chemical species, such as W4,[1] HEAT,[2–4] FPD,[5] ANL1,[6] some of the explicitly correlated approaches,[7] *etc.* Since thermochemistry essentially provides information for the stationary points on a potential-energy surface, ATcT can certainly provide at least a partial validation for ML potential models. In order to give just one example as an illustration, the pairwise interaction potential of water molecules essentially corresponds to the potential-energy surface of water dimer, and one of the elementary requirements should be that it correctly reproduces the dimerization energy of water.[8] Of course, further detailed validation may involve solving the ML potential for the rovibrational levels and comparing them to observed transitions using a spectroscopic database. An additional validation might involve the analysis of the attractive portion of the ML potential at long ranges in terms of dipole–dipole, dipole–quadrupole, *etc.* interactions and compare the results to independent expectations for an incipient hydrogen bond. Similarly, an analysis of the repulsive portion of the ML potential should correctly reproduce the 'excluded volume' term (which is, together with the dimer equilibrium constant, a constituent of the second virial coefficient) in the equation of state of steam, *etc.* Finally, when used in simulations of bulk liquid water and ice, the ML potential should correctly reproduce the thermal dependence of various water properties, including the density, where until very recently we were in the situation that models insisted that the Titanic disaster could not be blamed on a collision with an iceberg, since ice – according to modelling results – sinks.

1 A. Karton, E. Rabinovich, J. M. L. Martin and B. Ruscic, W4 theory for computational thermochemistry: In pursuit of confident sub-kJ/mol predictions, *J. Chem. Phys.*, 2006, **125**, 144108, DOI: 10.1063/1.2348881.
2 M. E. Harding, J. Vázquez, B. Ruscic, A. K. Wilson, J. Gauss and J. F. Stanton, High-accuracy extrapolated ab initio thermochemistry. III. Additional improvements and overview, *J. Chem. Phys.*, 2008, **128**, 114111, DOI: 10.1063/1.2835612.
3 J. H. Thorpe, C. A. Lopez, T. L. Nguyen, J. H. Baraban, D. H. Bross, B. Ruscic and J. F. Stanton, High-accuracy extrapolated *ab initio* thermochemistry. IV. A modified recipe for computational efficiency, *J. Chem. Phys.*, 2019, **150**, 224102, DOI: 10.1063/1.5095937.
4 J. H. Thorpe, J. L. Kilburn, D. Feller, P. B. Changala, D. H. Bross, B. Ruscic and J. F. Stanton, Elaborated thermochemical treatment of HF, CO, N_2, and H_2O: Insight into HEAT and its extensions, *J. Chem. Phys.*, 2021, **155**, 184109, DOI: 10.1063/5.0069322.
5 D. Feller, K. A. Peterson and B. Ruscic, Improved accuracy benchmarks of small molecules using correlation consistent basis sets, *Theor. Chem. Acc.*, 2014, **133**, 1407, DOI: 10.1007/s00214-013-1407-z.
6 S. J. Klippenstein, L. B. Harding and B. Ruscic, Ab Initio Computations and Active Thermochemical Tables Hand in Hand: Heats of Formation of Core Combustion Species, *J. Phys. Chem. A*, 2017, **121**, 6580–6602, DOI: 10.1021/acs.jpca.7b05945.

7 W. Klopper, B. Ruscic, D. P. Tew, F. A. Bischoff and S. Wolfsegger, Atomization energies from coupled-cluster calculations augmented with explicitly-correlated perturbation theory, *Chem. Phys.*, 2009, **356**, 14–24, DOI: **10.1016/j.chemphys.2008.11.013**.
8 B. Ruscic, Active Thermochemical Tables: Water and Water Dimer, *J. Phys. Chem. A*, 2013, **117**(46), 11940–11953, DOI: **10.1021/jp403197t**.

Fernanda Duarte said: You mentioned that the ATcT website is accessed every 2 seconds. What about data generation; who is generating the data? We have mentioned data is needed but tend to outsource this task.

Branko Ruscic answered: Indeed, we are receiving a million page-views each month, which translates on average to one access every 2.5 seconds. In order to answer your question on generating data, perhaps the best thing to do is to mention the various stages in reverse, starting from the website. The pages on the website are generated dynamically by pulling the necessary information from the SQL database. The website also possesses an archival quality similar to printed journals, allowing permanent access to all previous versions of results. When a new version of ATcT data is ready for publication on the website, it gets transcoded and uploaded to the SQL database on the public ATcT server by dedicated software written in our group.

Generating a new version of ATcT data, however, does involve a substantial amount of curation by the ATcT team, which starts by identifying additional chemical species of interest, performing literature searches for each species, critical evaluation of the relevant papers and extraction of the actually determined quantities, passing judgement on the initial uncertainty, performing – if necessary – additional electronic structure calculations (by our team, by the extended team organized as an ATcT Task Force, or by other external collaborators) or additional experimental measurements (usually in collaboration with appropriate experts), and then exercising the ATcT software to construct the thermochemical network, to statistically analyze each determination against the cumulative knowledge contained in the network, and to solve the network simultaneously for all species. There are also post-processing actions, such as checking the log of the ATcT run for potential problems, examining the list of outliers that ATcT found during the run, performing variance decomposition for each species in order to obtain the provenance of each ATcT value, *etc.* If problems are found, then the ATcT run needs to be repeated after appropriate corrective action. The current aim is to annually publish on the ATcT website a couple of new (augmented) versions of ATcT results and keep fulfilling the role expected of a DOE Office of Science Public Reusable Data Resource.

Niamh Hickey asked: Is there a way of streamlining the information in this database into different chemical categories so it is easier to search the database?

Branko Ruscic answered: We are currently testing API calls that will enable automated retrieval of the ATcT data as a JSON object, ranging from retrieval of the thermochemistry of one or more individual species or of stoichiometrically balanced chemical reactions, up to and including the retrieval of the complete set of data in the current ATcT version. The ATcT data is currently accompanied with rather rich metadata, including a variety of species identifiers (such as chemical formula, molecular weight, alternate chemical names, CAS registry numbers,

InChI strings, SMILES, *xyz* geometries in the Wigner inertial frame, *etc.*). Depending on what the desired classification categories are, these may already be enough, or may need to be combined with additional descriptors (from external databases or tools that can generate them) that will in turn help achieve the desired classification into user-designated categories.

Filip T. Szczypiński commented: You have analysed an immense amount of datasets originating from multiple sources. As a physical organic chemist, I know that the measured values very often depend on the exact protocol that is followed by the researcher and that different research groups adapt slight variations to the same experimental techniques. With such a wide systematic analysis of data coming from across the scientific community, are you planning to extract the most reliable approaches to physical measurements and – if so – would you be interested in publishing a series of guidelines on different analytical techniques to help perplexed experimentalists and thus improve the quality of the data that we generate?

Branko Ruscic replied: That sounds like an excellent suggestion, and the answer is affirmative, at least in principle. Of course, writing such guides requires detailed expertise in the targeted area, as well as considerable experience, in order to avoid succumbing to the detrimental Dunning–Kruger effect. We have, in fact, already done something along the lines of your suggestion. For example, some time ago we produced a practical guide on how to correctly extract from photoionization mass spectrometric measurements a fragment appearance energy and how to correctly determine the adiabatic ionization energy of the corresponding transient species, with the aim of closing the 'positive ion cycle' and producing a rather accurate experimental bond dissociation energy.[1] The guide has been subsequently perused by researchers at several synchrotron sources who were awarded the beamtime but lacked the necessary background experience in laboratory photoionization. More recently, we have composed a brief guide for theorists explaining the convention of expressing the uncertainty of thermochemical quantities, together with tips on how to benchmark computed thermochemical quantities.[2] Finally, our group has recently put together a practical guide for thermochemistry, with the aim of explaining, *inter alia*, a number of current thermochemical conventions, the importance of anharmonic effects, the details of the formats of NASA (and other) thermochemical polynomials, *etc.*[3] We certainly have the best intention to continue these efforts.

1 B. Ruscic, Photoionization Mass Spectroscopic Studies of Free Radicals in Gas Phase: Why and How, in *Research Advances in Physical Chemistry*, ed. R. M. Mohan, Global, Trivandrum, India, 2000, vol. 1, pp. 39–75.
2 B. Ruscic, Uncertainty quantification in thermochemistry, benchmarking electronic structure computations, and active thermochemical tables, *Int. J. Quantum Chem.*, 2014, **114**, 1097–1101, DOI: **10.1002/qua.24605**.
3 B. Ruscic and D. H. Bross, Thermochemistry, in *Mathematical Modelling of Gas-Phase Complex Reaction Systems: Pyrolysis and Combustion*, ed. T. Faravelli, F. Manenti and E. Ranzi, Computer Aided Chemical Engineering, Elsevier, 2019, ch. 1, vol. 45, pp. 3–114, DOI: **10.1016/B978-0-444-64087-1.00001-2**.

Christian Kuttner communicated: Uncertainties indicate the degree of confidence we can have in a specific piece of data or information. Could you clarify the

different types of uncertainties that are commonly encountered? Based on your experience and perspective, what should be the recommended standards for reporting uncertainties in scientific literature?

Branko Ruscic communicated in reply: Indeed, a paper that reports a numerical value for a physical quantity without a clear indication of what is its expected accuracy is significantly less useful than a paper that includes some estimate of the accompanying uncertainty. As mentioned, the convention of expressing uncertainty for thermochemical quantities is a 95% confidence interval and has to include an earnest estimate of both random and all potential systematic sources of error. This convention is strictly followed by virtually all thermochemical tabulations. The underlying history is given elsewhere.[1,2] Notably, although scientists who use the tabulated thermochemical values are not necessarily aware that these are accompanied by 95% confidence intervals, they fully expect that the listed uncertainties can be directly compared across different sources, allowing them, for example, to select the value that seems most accurate. Of course, in order to be directly comparable, the uncertainties need to have the same meaning, or 'coverage factor'. The required 95% confidence interval has a nominal coverage factor of 2. One particular problem was introduced by theorists who compute thermochemical quantities using electronic structure methods, but express the uncertainty *via* the benchmarked mean absolute deviation (MAD) of the theoretical method. MAD is, in fact, smaller than one standard deviation – having a coverage factor that is dependent on the distribution, but is usually $\sim$0.7 or even less and at most 0.8 – and thus underestimates the expected uncertainty by a factor of $\sim$3 or more. Similarly, expressing the uncertainty *via* the benchmarked root mean square error, RMSE, of the theoretical method, which essentially has a coverage factor of 1, underestimates the expected uncertainty by a factor of 2 (or more, depending on the size of the benchmark set). In addition, a few theorists use rather bizarre variants of the above, such as expressing the uncertainty as a mean (signed) deviation, followed, after a $\pm$ sign, by a measure of the spread of the deviations around their mean (where the latter can be either a standard deviation or a 95% confidence interval). This representation can be quite misleading and confusing to the reader. The quoted spread essentially says something about the precision, rather than the accuracy (see ref. 2 for a definition of accuracy, precision, and trueness). In these cases, one essentially needs to either completely redo the benchmarking process, or very approximately estimate the uncertainty by adding up the absolute values of the mean deviation and its spread and then multiply by the appropriate coverage factor.

In many other areas of physical chemistry and chemical physics the standards for expressing uncertainties seem to be less firmly established. Spectroscopists frequently use standard deviations and tend to present them at the tail of the reported numerical value as additional digits in parentheses. Many kineticists do not provide an indication of uncertainty at all for their measured or computed kinetic rate constants, and when they do, these are most frequently standard deviations and occasionally 95% confidence intervals. Another interesting aspect of uncertainties occurs in cases of fitted parameters that are in reality highly correlated. Two examples of the above group that come to mind are the expression of a kinetic rate constant in terms of fitted values for the prefactor A and activation energy E_a of the Arrhenius expression, $k_T = A\exp(-E_a/(kT)$, or the determination

of the reaction enthalpy $\Delta_r H_T^\circ$ and entropy $\Delta_r S_T^\circ$, by performing a van't Hoff fit of an equilibrium constant (itself obtained, for example, by pairing forward and reverse kinetic rate constants or by measuring equilibrium concentrations). Although sometimes the researchers provide uncertainties for both fitted parameters, they virtually never provide their covariance or correlation coefficient, making it impossible to correctly reconstruct the actual uncertainty of their fitted rate constant or equilibrium constant.

The situation of missing, nonstandard, or poorly defined uncertainties repeats in nearly all other areas of science. A welcome move toward significantly improving the current situation would occur if scientific and technical journals (as well as data depositories) would start strictly requiring that reported (or deposited) numerical results be accompanied by uncertainties that follow the standards in the field, or, in the absence of established standards, explicitly define the nature of the reported uncertainties. Admittedly, many data generators will perceive such a requirement as an additional burden because estimating uncertainties can sometimes be a rather difficult task. However, the ISO Guide to Expression of Uncertainty in Measurements (ISO GUM)[3] provides a detailed guidance on how to evaluate and combine Type A (obtained by benchmarking) and Type B (estimated from experience) uncertainties. NIST has a simpler, easy to follow guide on estimating both type of uncertainties.[4]

1 B. Ruscic, Uncertainty quantification in thermochemistry, benchmarking electronic structure computations, and active thermochemical tables, *Int. J. Quantum Chem.*, 2014, **114**, 1097–1101, DOI: 10.1002/qua.24605.
2 B. Ruscic and D. H. Bross, Thermochemistry, in *Mathematical Modelling of Gas-Phase Complex Reaction Systems: Pyrolysis and Combustion*, ed. T. Faravelli, F. Manenti and E. Ranzi, Computer Aided Chemical Engineering, Elsevier, 2019, ch. 1, vol. 45, pp. 3–114, DOI: 10.1016/B978-0-444-64087-1.00001-2.
3 Evaluation of Measurement Data — Guide to Expression of Uncertainty in Measurement, JCGM 100:2008(E), Joint Committee for Guides in Metrology: Sevres, 2010.
4 B. N. Taylor and C. E. Kuyatt, *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, NIST Technical Note 1297, U.S. Gov. Printing Office: Washington, DC, 1994.

Wenhao Sun opened a general discussion: Alán mentioned earlier that we should consider the synthesizability of compounds while we do these property screens, and that we should do this by considering convex hull stability. I think in recent years we are increasingly realizing that convex hull stability is not a very good metric of synthesizability. Many convex hull metastable compounds are readily synthesizable and in use in functional devices; meanwhile, there are many convex hull stable phases that cannot be readily synthesized (their formation gets 'blocked' by the formation of metastable intermediates). These are not issues regarding DFT formation energy errors (even though that is an issue)—this is a fundamental lack of synthesis science regarding the synthesizability of metastable materials. More importantly, convex hull stability does not provide any actionable guidance on how to actually synthesize a material experimentally. A pragmatic theory for materials synthesis should aim to answer three tiers of questions: (i) Is a predicted material synthesizable? (ii) If yes, what synthesis method – *e.g.* solid-state, hydrothermal, vapor deposition, *etc.* – is best to synthesize it? (iii) Within the parameter space of that synthesis method, what 'recipe' can lead to a phase-pure synthesis? If we want to talk about predictive synthesis, I think we need to answer all 3 tiers of the synthesizability question. Otherwise I think our predictions will have very limited utility to experimentalists.

For metastable materials and their synthesizability, refer to ref. 1.

For stable materials that are difficult to synthesize, refer to ref. 2 and 3. Dolomite is a very stable mineral that forms mountains in nature but is impossible to grow in the laboratory.[2] The ScZn quasicrystal is convex-hull stable but is nucleation-limited across the phase diagram.[3]

1 W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson and G. Ceder, The thermodynamic scale of inorganic crystalline metastability, *Sci. Adv.*, 2016, **2**(11), e1600225, DOI: **10.1126/sciadv.1600225**.
2 J. Kim, Y. Kimura, B. Puchala, T. Yamazaki, U. Becker and W. Sun, Dissolution enables dolomite crystal growth near ambient conditions, *Science*, 2023, **382**(6673), 915–920, DOI: **10.1126/science.adi3690**.
3 W. Baek, S. Das, S. Tan, V. Gavini and W. Sun, Quasicrystal bulk and surface energies from density functional theory, *arXiv*, 2024, preprint, arXiv:2404.05200, DOI: **10.48550/arXiv.2404.05200**.

Graeme M. Day answered: I agree with Wenhao Sun that convex hull stability does not provide a complete picture of the synthesizability of a material. If we exclude materials that are above the convex hull, we could miss some materials that can be experimentally accessed under the right conditions. Over the past few years in the field of organic molecular crystals, we have found that crystal structures that are predicted to lie well above the most stable structure for a given composition can have exceptional properties and can be made in the lab.[1,2] Our problem is that many predicted structures appear at similar energies to these synthesizable structures. So we need other calculable ways to distinguish those that are unsynthesizable from those that can be made in the lab. A method that we have found to be promising is to calculate the barriers around local energy minima: high-energy structures that we find experimentally seem to correspond to structures in deep energy wells, so that transformation to the globally more stable structures is kinetically hindered.[3]

I don't agree that predictions are not valuable if we cannot predict the synthesis method. Predictions that tell us where to look (*e.g.* which molecule, or which composition) are of utility in guiding experimental programmes, where a range of conditions can then be applied to attempt to synthesize promising predicted structures. Automation in the lab will help us perform screens over a range of conditions, particularly as robots are able to handle solids and perform crystallisation experiments.[4,5] Predicting synthesis conditions will be valuable if it can be done, but current methods have been demonstrated to be useful without predicting synthesis conditions for predicted materials.

1 A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper and G. M. Day, Functional materials discovery using energy–structure–function maps, *Nature*, 2017, **543**, 657–664, DOI: **10.1038/nature21419**.
2 C. M. Aitchison, C. M. Kane, D. P. McMahon, P. R. Spackman, A. Pulido, X. Wang, L. Wilbraham, L. Chen, R. Clowes, M. A. Zwijnenburg, R. S. Sprick, M. A. Little, G. M. Day and Andrew I. Cooper, Photocatalytic proton reduction by a computationally identified, molecular hydrogen-bonded framework, *J. Mater. Chem. A*, 2020, **8**, 7158–7170, DOI: **10.1039/d0ta00219d**.
3 S. Yang and G. M. Day, Global analysis of the energy landscapes of molecular crystal structures by applying the threshold algorithm, *Commun. Chem.*, 2022, **5**, 86, DOI: **10.1038/s42004-022-00705-4**.
4 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng and G.

Ceder, An autonomous laboratory for the accelerated synthesis of novel materials, *Nature*, 2023, **624**, 86–91, DOI: 10.1038/s41586-023-06734-w.
5 A. M. Lunt, H. Fakhruldeen, G. Pizzuto, L. Longley, A. White, N. Rankin, R. Clowes, B. Alston, L. Gigli, G. M. Day, A. I. Cooper and S. Y. Chong, Modular, multi-robot integration of laboratories: an autonomous workflow for solid-state chemistry, *Chem. Sci.*, 2024, **15**, 2456–2463, DOI: 10.1039/d3sc06206f.

Jeremy Frey remarked: Synthesis is a challenge. Using the available literature knowledge, machine learning techniques can provide predictive models based on current synthetic approaches, which then lays down a challenge to colleagues involved in synthesis to develop new techniques to synthesize molecules that challenge current approaches. Many years ago I read a book titled "Non-existent compounds".[1] When I was reading it many had in fact been made; a testament to the ingenuity of synthetic chemists.

1 W. E. Dasent, *Nonexistent Compounds: Compounds of Low Stability*, Marcel Dekker, Inc., New York, 1965.

Steven Torrisi addressed: I'd like to make a comment about synthesizability to add to the discussion with the earlier points made by Alán Aspuru-Guzik and Wenhao Sun: thermodynamics-based measures like the convex hull stability are very attractive to computationalists (like myself!), because it provides a simple basis for comparison, it makes use of the databases we have, and the idea has a firm principled basis. But I fear that there is sometimes a 'computability bias' that we have, as a community – a little like the streetlight effect.[1] In summary, the story goes of a drunk man looking for his lost keys in the street, when he knows he lost them in the bushes, because he's searching where the light is. The analogy here is studying simplified pictures of inorganic solid-state materials because they are far easier to handle (and critically, are computationally tractable in DFT), than models which incorporate, *e.g.*, defected structures, larger-scale structures, and so on. Not only are our tools best suited towards small unit cells, but also, we have some empirical evidence from Lucas Wagner at UIUC from 2016,[2] and recent work from Toyota Research Institute in 2024,[3] that even thermodynamically favorable structures can be challenging to synthesize. This comment is not intended as a critique of the practice of using the hull energies—nor the great work by other delegates in complicating that picture to improve our understanding and predictive power of synthesis. It is merely to highlight that our methods of abstraction have some intrinsic limitations owing to necessary simplification.

1 https://en.wikipedia.org/wiki/Streetlight_effect.
2 A. Narayan, A. Bhutani, S. Rubeck, J. N. Eckstein, D. P. Shoemaker and L. K. Wagner, Computational and experimental investigation for new transition metal selenides and sulfides: The importance of experimental verification for stability, *Phys. Rev. B*, **94**, 045105, DOI: 10.1103/physrevb.94.045105
3 J. H. Montoya, C. Grimley, M. Aykol, C. Ophus, H. Sternlicht, B. H. Savitzky, A. M. Minor, S. B. Torrisi, J. Goedjen, C.-C. Chung, A. H. Comstock and S. Sun, How the AI-assisted discovery and synthesis of a ternary oxide highlights capability gaps in materials science, *Chem. Sci.*, 2024, **15**, 5660–5673, DOI: 10.1039/d3sc04823c.

Christopher M. Collins replied: I would like to add to this debate, that it is entirely possible to embrace this observation; that due to the required granularity of computational searches, along with the great difficulty of predicting disorder, it is difficult to imagine that we would ever be able to exactly predict the

composition and structure of new compounds. So in our work, we don't. Instead, we consider the structures which we generate "probe structures", which we define as hypothetical structures, which contain local co-ordinations which are representative of potential single phases which could form, but we do not explicitly worry about whether or not they are the ground state. The structure prediction methods that I have developed (MC-EMMA and FUSE), are then configured with this in mind. We can use these probe structures to predict an upper bound on the energy *vs.* a convex hull – since we know that the true ground state of the composition will either be the energy of the probe structure, or lower. We use this to map chemical space for regions, rather than specific compositions with low energies; in these regions, we then can target our synthetic efforts.

We now have several examples of using this approach, combined with experimental exploration at and around the compositions of probe structures, which has led directly to the synthesis of new compounds. Two notable examples of these are ref. 1 and 2. In the first case, we arrive at two oxide materials with complex, disordered structures, which we would never expect a structure prediction method to be able to predict. In the second example, we also combine probe structure prediction with composition-only properties prediction, since as we assume that we will not find the true ground state structure, it follows that it makes no sense to predict properties using predicted structures! This project resulted in the discovery of an oxide with an extremely low thermal conductivity (which was our target property); at the time it was the lowest for any oxide containing a first-row transition metal. The structure of this compound was also the first bulk oxide quasi-crystal, where the nearest commensurate approximant model would contain many millions of atoms, which obviously, we would never expect a crystal structure prediction algorithm to predict.

1 C. Collins, M. S. Dyer, M. J. Pitcher, G. F. S. Whitehead, M. Zanella, P. Mandal, J. B. Claridge, G. R. Darling and M. J. Rosseinsky, Accelerated discovery of two crystal structure types in a complex inorganic phase field, *Nature*, 2017, **546**, 280–284, DOI: **10.1038/nature22374**.

2 C. M. Collins, L. M. Daniels, Q. Gibson, M. W. Gaultois, M. Moran, R. Feetham, M. J. Pitcher, M. S. Dyer, C. Delacotte, M. Zanella, C. A. Murray, G. Glodan, O. Pérez, D. Pelloquin, T. D. Manning, J. Alaria, G. R. Darling, J. B. Claridge and M. J. Rosseinsky, Discovery of a Low Thermal Conductivity Oxide Guided by ProbeStructure Prediction and Machine Learning, *Angew. Chem. Int. Ed.*, 2021, **60**, 16457–16465, DOI: **10.1002/anie.202102073**.

James Proudfoot addressed Kim E. Jelfs, Arya Changiarath and Yuchen Lou: For Bayesian optimisation, what are your opinions on the choice of initialisation and the choice of termination conditions? How do you select the initial pool for Bayesian optimisation? How do you know when to stop Bayesian optimisation – a fixed number of steps or a specific target property value being looked for?

Kim E. Jelfs responded: Of course this is very case study specific and within WebBO there are no specific choices predetermined and it is up to the user to make these decisions. We have seen that the choice of how to initialise can heavily influence the result, and so it is important for a user to be mindful of this. With regards to when to stop, of course, this can be resource dependent, when the time/chemicals/cost becomes prohibitive, or the user could decide a termination criteria related to performance; for example, when a yield of 99% or above is reached.

Arya Changiarath responded: The initial pool of sequences contains the diverse set of sequences from fine-tuned ProteinGPT and standard protein GPT with a range of B_{22} values corresponding to sequences that could phase separate, and ones that were not able to phase separate. This ensures that the optimisation explores the broad range of parameter space. We trained the model until its predictions matched the calculated values from the simulations in the correlation plot. At the same time, we also tracked the numeric labels for the validation set sequences and measured the mean squared error (MSE) between the true values from the simulations and the model's predicted values at each optimization step.

Yuchen Lou replied: We referenced the hyperparameter choice from ref. 1. The initial choice range was relatively wide and we stopped Bayesian optimization mainly on the concern of time. Once the run seemed to have explored sufficient parameter space, we followed up with grid-based optimisation close to the optimal parameters identified from BO. We felt like the run had achieved enough exploration of the entire range and that doing a simple grid tuning was the best exploitation rather than continuing BO.

1 Z. Chen, N. Andrejevic, T. Smidt, Z. Ding, Q. Xu, Y.-T. Chi, Q. T. Nguyen, A. Alatas, J. Kong and M. Li, Direct Prediction of Phonon Density of States With Euclidean Neural Networks, *Adv. Sci.*, 2021, **8**, 2004214, DOI: **10.1002/advs.202004214**.

Andrew I. Cooper said: If you consider the huge diversity of chemistry, does it always work trying to build descriptors? That is – is it always sensible to try to try to be 'driven' by computational data, which is a main theme of this meeting? For example, there are lots of mesoscale systems where it is challenging to build computational descriptors. Is it then better to use experimentation and to look to maximise the outputs (function) rather than the inputs (descriptors)? In essence, are there cases where one might expend a lot of energy and actually go slower with 'data driven' methods?

Jörg Saßmannshausen addressed Heather J. Kulik, Daniel Crusius, Sanggyu Chong and Branko Ruscic: I have a general question: What is it that the synthetic community needs to do when uploading raw data to sites like Zenodo, Figshare, *etc.?* What would be the best data format? Should failed experiments be included as well? What kind of metadata is valuable and do we have some kind of ontology to fall back on?

Heather J. Kulik replied: One of the biggest challenges we encounter is that a lot of useful information is stuck in supporting information. I definitely think that data in comma delimited files or other database formats would be an awesome step forward, including digitized graph data. Nevertheless, there are major advances being made in graph digitization that might reduce the need for that. I can't say what the best data format would be – although I know in the context of organic reactions, folks have come up with a schema to encourage uniform data collection. And absolutely, failed experiments as well as variability based on sample (*e.g.*, repeated measurements) would be great. Metadata that is valuable is reaction conditions, temperature, *etc.* I think though the right thing is to have experimental communities come together and agree upon a standard – or

at least it should come from an authority who understands the challenges associated with collecting the measurements in the first place.

Daniel Crusius answered: Failed experiments should always be included, ideally also repeat measurements to get an estimate of experimental errors. The best data format will likely depend on the specific type of experiment performed, and best practices, including which metadata to add, will need to be established by the respective communities.

Sanggyu Chong replied: To add my own perspective here—recognizing that such data needs to be crawled and mined by the data scientists on the other end, it would be immensely beneficial for the experimentalists to spend a little bit of time familiarizing themselves with how their data will be processed by the data scientists, and liaise with them if possible. Ideally, experimentalists should remain at the disposal of anyone in need of additional information on their data, but I recognize that this is far too utopian.

Branko Ruscic added: Clearly, deposited data needs to be accompanied not only by uncertainties that indicate the expected accuracies of each item, but also by abundant metadata. The metadata needs to be sufficiently rich to unambiguously define the nature and character of the data, its provenance, and include as many useful descriptors as practical. Of course, the actual details of what type of metadata is needed or desired will depend on which area of science the deposited data belongs to. The actual format of the data is probably less important, as long as it is machine readable and sufficiently decorated with metadata, since the user can always write a few lines of script that will translate the deposited data to the desired format.

Mark Goulding said: Branko raised the importance of data provenance in published datasets, in order for the data to be reusable with confidence. I would like to add a comment to this.

Within industry, where many decisions are made based on accurate structure/property datasets, a significant amount of work over long time periods is made to ensure high data integrity:

(i) Test methods are standardised and defined by standard operating procedure.

(ii) Equipment and instruments are regularly calibrated.

(iii) Dedicated screening labs with experienced personnel may be established.

(iv) A stable process of tests, equipment and labs may exist for several years.

All of this leads to data with high quality and consistency, but there may be significant cost to establish and maintain the process. Within industry this can often be justified.

It is also a key reason that publication of such datasets is seldom made and the data remains proprietary. Often it is the life blood of future profitable business.

Branko Ruscic replied: I would entirely agree with your comment, and add that – of course – doing research at universities and national laboratories also carries a substantial cost, probably not very different from that in industry. In fact, as opposed to the largely secretive cost in case of industry, one can easily estimate the average cost of a scientific paper at universities or national laboratories simply by

dividing the amount of the grant by the total output in terms of papers, typically arriving at costs of the order of one or more hundred thousand dollars per paper.

It should be also mentioned that in most areas of physical chemistry it is rather clear how the experiments need to be conducted in order to produce results that are physically meaningful (which is an almost universal requirement for scientific data). The stipulations followed in industrial work that you have mentioned equally apply to academia. The main difference from industry is that the bulk of research at universities and national laboratories is funded by taxpayers, and thus the related results are generally expected to be published in scientific journals, reports, or deposited in various publicly accessible depositories, driving the "publish or perish" paradigm in academia. Of course, there are numerous cases to the contrary in the described academia *vs.* industry dichotomy in data openness or secrecy. For example, when industry hires research groups at universities or national laboratories to perform some research for them, the results in most cases remain the property of the private funder even though the research has been conducted in academic environments. Also, some research at universities and national laboratories may be or become classified for reasons of national security. On the other hand, there are numerous cases when researchers in industry are encouraged to publish at least some of the findings in a scientific journal (notorious examples would be the original Bell Labs, the now defunct Amoco Labs, *etc.*). There are also cases where researchers in industry surreptitiously publish their findings. An example of the latter that immediately comes to mind is William S. Gosset, the English statistician and chemist who was the head brewer at Guinness, but published profusely under the pseudonym "Student", famously becoming the eponym for the Student-*t* test.[1]

1 Student, The Probable Error of a Mean, *Biometrika*, 1908, **6**, 1–25, DOI: **10.2307/2331554**.

Christian Kuttner responded: I fully agree with the points raised regarding the importance of data integrity in both industry and academic settings. Proper training, as highlighted above, is indeed critical, and it complements the standardized procedures and equipment calibration in ensuring reliable, reproducible data. These elements together contribute significantly to the overall quality of scientific work, both in industry and academia.

Jörg Saßmannshausen replied: Having worked in industry, I would like to add an important point here:

(v) Proper training. Having received very good training as a lab-assistant of chemistry, I do value proper training very much. The proper use of lab equipment like pipettes, scales, syringes, equipment, *etc.*, proper lab-book keeping and recording all observations, and any mistakes, is very important. I was not always convinced this is done well in some universities to be frank. In the end, if a teacher does not know any better, how can the student do something well?

Rob Evans communicated: There was a lot of discussion at the end of the session about the suitability and quality of uploaded data. Individual scientists have their own individual practices which can evolve into best practices. The recent emergence of protocols and/or methods papers strikes me as an opportunity for research communities to develop and disseminate their own best practices, details as to the appropriateness of preparation and presentation of

data, as well as the language used in describing experiments, results, data, *etc.* There's a journal in the NMR field that employs a continuous, open peer-review model, which can flag up very interesting insights and comments.

Christian Kuttner communicated in reply: You're absolutely right that individual practices, when shared and refined, can lead to broader best practices that benefit the entire research community. The rise of protocols and methods papers offers a valuable opportunity for researchers to collaboratively define what "best practice" looks like in their field, ensuring that data is prepared and presented in ways that enhance reproducibility and clarity.

An 'open review' system, as you mentioned, is indeed an interesting approach, but I see it more as a supplementary process rather than a substitute for traditional peer review. The idea of rendering a published article into a 'living document' is undoubtedly valuable but also challenging, given the diversity of opinions in the scientific community.

When concerns arise after publication, formal avenues like letters to the editor or 'matters arising' articles[1] offer structured opportunities for issues to be raised and addressed. These mechanisms allow for both the concern and the author's response to undergo peer review, maintaining the integrity of the discourse. This process can lead to corrections, addenda, or even retractions when necessary, though it can often become quite complex. Still, it ensures that the scientific record remains robust and reliable, while allowing for the natural evolution of ideas.

1 **https://www.nature.com/ncomms/submit/matters-arising**.

Conflicts of interest

Matthew R. Ryder: This manuscript has been co-authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains, and the publisher, by accepting the article for publication, acknowledges that the US government retains, a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (**http://energy.gov/downloads/doe-public-access-plan**). Christian Kuttner is affiliated with Springer Nature as an editor for *Nature Communications*. The views expressed are their own and do not necessarily reflect the positions of *Nature Communications*, the Nature Portfolio, or Springer Nature. Jakob Zeitler declares the following competing financial interest(s): his work has included developing and commercialising Bayesian optimisation algorithms for Matterhorn Studio. There are no other conflicts to declare.

