Faraday Discussions
Cite this: *Faraday Discuss.*, 2025, **256**, 639

DISCUSSIONS

View Article Online
View Journal | View Issue

Discovering synthesis targets: general discussion

Andy S. Anker, Alán Aspuru-Guzik, Tim Bechtel, Filippo Bigi, Ksenia R. Briling, Basita Das, Nicholas David, Graeme M. Day, Volker L. Deringer, Matthew Dyer, Annabel Eardley-Brunt, Matthew L. Evans, Rob Evans, Barnabas A. Franklin, Alex M. Ganose, Janine George, Mark Goulding, Niamh Hickey, Gillian James, Adarsh V. Kalikadien, Venkat Kapil, Heather J. Kulik, Vishank Kumar, Christian Kuttner, Erwin Lam, Magdalena Lederbauer, Yuchen Lou, Jennie Martin, Andres Marulanda Bran, Miriam Mathea, Chris J. Pickard, Branko Ruscic, Matthew R. Ryder, Victor Sabanza Gil, Philippe Schwaller, Marwin H. S. Segler, Wenhao Sun, Sara Tanovic, Wojtek Treyde, Aron Walsh and Ruiqi Wu

DOI: 10.1039/d4fd90064b

Ruiqi Wu opened a discussion of the paper by Miriam Mathea: Could you specify the dimension of the fingerprints? Are they effective in identifying functional groups as clusters beyond the similarities on the atomic scale?

Miriam Mathea answered: Thank you very much for the question! The Morgan fingerprint consisted of 2048 features, encoding integer values, whereas the neural fingerprint consisted of 300 float values. Both fingerprints capture the topology from the molecular graph and therefore can capture molecular substructures like functional groups. However, the fingerprints capture substructures in different ways. The algorithm of Morgan fingerprints works by iteratively expanding from each atom in a molecule, generating a series of concentric circular subgraphs. Unlike classical fingerprints, graph neural networks have the ability to learn and encode structural information in a task-specific manner using message passing.

Filippo Bigi commented: You said that the accuracy of your model degrades when you change its final part. Have you tried some post-processing uncertainty approaches that don't change the predictions of the model?

Miriam Mathea replied: Thank you very much for the question! The scope of this project was to evaluate models and techniques frequently used in cheminformatics, and hence, Chemprop was calibrated only with methods readily

available in scikit-learn. We looked into other methods but integrating them into scikit-learn is out of scope, as the main objective was to assess the impact of neural fingerprints on default models.

Alán Aspuru-Guzik said: Great talk. I love your approach.

Have you considered using and benchmarking against the Laplace approximation?[1] It is readily available using a "one-line-of-code" approach in Pytorch.[2]

1 A. Kristiadi, M. Hein and P. Hennig, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, 2021, **161**, 344–353, **https://proceedings.mlr.press/v161/kristiadi21a.html**.

2 E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer and P. Hennig, *arXiv*, 2021, preprint, arXiv:2106.14806, DOI: **10.48550/arXiv.2106.14806**.

Miriam Mathea responded: Thank you very much for the suggestion! We decided not to implement this model, as the scope of this project was to use models frequently used in cheminformatics and evaluate the impact of the neural fingerprint compared to the standard Morgan fingerprint. However, this approach seems promising, and we concluded that a follow-up study would be relevant, assessing a broader spectrum of algorithms to determine best practices for model calibration.

Philippe Schwaller asked: For work in industrial settings, how are the uncertainties used/shown to chemists?

Miriam Mathea answered: In the context of decision-making in the field of chemistry, it is essential for chemists to have access to relevant information that can aid in their decision-making process. Providing a concise overview of the results along with additional contextual information can be valuable in assessing the reliability of predictions.

One approach is to display the predicted probability alongside the predicted label. This allows chemists to have a sense of confidence in the model's prediction. The applicability domain refers to the region in which the model's predictions can be considered reliable. By defining and visualizing the applicability domain, chemists can assess whether a specific prediction falls within the reliable range of the model. This can be done by, *e.g.*, presenting the chemists with the nearest neighbors from the training set. This allows chemists to compare the predicted compound with similar compounds from the training set, providing additional context and aiding in the decision-making process.

Additionally, a 2D mapping of the chemical space of the training set can be provided and explainable AI methods can support interpretation of results.

Vishank Kumar remarked: By including fingerprints, the accuracy is reduced and uncertainty gets better. Why?

Miriam Mathea replied: For standard ML models, the accuracy was not reduced by using the neural fingerprint, but the calibration reduced the performance of Chemprop. We assume that the imbalance in the data was not fully captured by the calibration method, shifting the decision boundaries.

Vishank Kumar queried: Have you taken the model/uncertainty estimates to your colleagues who are synthesising materials and applied it? What were the learnings from such an exchange? Additionally, how do you think we can narrow this gap between computational predictions and the synthesis lab?

Miriam Mathea responded: We have had some discussions with colleagues involved in material synthesis regarding the model and uncertainty estimates. The exchange provided a general understanding of their perspectives and some insights into the practical challenges they face. Moving forward, fostering ongoing communication and collaboration could help bridge the gap between computational predictions and the synthesis lab, allowing for a more integrated approach to our work.

Basita Das enquired: Have you tried using two different types of experimental data to reduce uncertainty?

Miriam Mathea answered: No, we have not attempted to use two different types of experimental data to reduce uncertainty because the underlying dataset only contains experimental data from one source. Future analysis of uncertainty could benefit from including a dataset that consists of multiple sources of experimental data.

Annabel Eardley-Brunt communicated: In Fig. 1 of your paper (**https://doi.org/10.1039/d4fd00095a**), the balanced accuracy error (shown in the error bars) is quite different for some model combinations compared to others. Why do you think this is the case?

Miriam Mathea communicated in reply: Thank you very much for the question! To the best of our knowledge, there is no known explanation for why some models have higher variability in balanced accuracy. We don't expect a systematic trend in the difference in the spread. We assume it comes from the nuances in featurization and methodology dependent on the data.

Christian Kuttner communicated: Balancing prediction accuracy and uncertainty is challenging. Your study (**https://doi.org/10.1039/d4fd00095a**) finds that combining neural fingerprints with classical ML methods provides better-calibrated uncertainty estimates, although there is a slight drop in prediction performance compared to the native Chemprop model. How can this balance between prediction accuracy and uncertainty be further optimized, especially for models applied in real-world chemical discovery?

Miriam Mathea communicated in reply: The idea to use the neural fingerprint with standard ML models originated from the observation that Chemprop was often not well calibrated. Exploring new prediction layers or loss functions could also represent viable solutions to this problem.

Christian Kuttner communicated: Your results suggest that neural fingerprint-based methods, particularly when used with random forests, are more robust in providing uncertainty estimates for molecules dissimilar to the training set. What

specific properties of neural fingerprints contribute to this robustness, and how can these properties be leveraged in other areas of molecular modeling?

Miriam Mathea communicated in reply: The difference between using neural fingerprints in contrast to Morgan fingerprints in this scenario is that the neural fingerprints are a learned representation with features extracted based on the target property. This learned representation might be more robust than the static substructures in the Morgan fingerprint in the scenario with more dissimilar molecules to the training set. Learned fingerprints are already applied and popular in many areas of molecular modeling.

Matthew R. Ryder communicated: Your comparison of neural fingerprints and traditional machine-learning methods highlights improvements in uncertainty estimation. How do you envision using uncertainty quantification to improve predictions in dynamic systems, such as those with significant disorder or non-covalent interactions?

Miriam Mathea communicated in reply: Our work focused on property prediction problems from the molecular graph. We did not consider molecules' conformations or dynamic systems. While we believe that investigating similar methods for dynamic systems might give interesting insights, these require a substantially different study because the molecular representation and data differ significantly.

Wojtek Treyde opened a discussion of the paper by Marwin H. S. Segler: In your opinion, what other applications of machine learning in chemistry currently lack such benchmarking and evaluation suites, and is addressing this something that you're actively working on?

Marwin H. S. Segler answered: Apart from protein crystal structure prediction, there is probably no other well-defined evaluation for any of the many other important prediction problems in chemistry. One example of a great initiative is Polaris, which addresses the problem of molecular property prediction.

Sara Tanovic commented: The top-k accuracies of different retrosynthesis models tend to be within a few percentage points of each other when tested on the same test set. Do you think there's a statistical limit to this metric that depends on the test set itself?

Marwin H. S. Segler replied: Very likely yes. Also, it has to be noted that some reactions are somewhat redundant (*e.g.*, cross couplings with different leaving groups).

Mark Goulding said: Please can you comment on the fact that if you're using USPTO data set sequences or a data set from published papers, those are likely to be a best case synthesis of a compound. They may not be as good as a dataset (*e.g.*, an entire synthesis dataset from an electronic laboratory notebook software application (ELN) that contains both the syntheses with a positive outcome and those with a negative outcome). How does this affect the models?

Marwin H. S. Segler responded: USPTO can be useful for testing, and for public benchmarking is probably the best there is; however, for production use it is recommended to use better curated in house datasets, including those from ELNs.

Philippe Schwaller asked: If you had to come up with a recommendation for a multi-step direct synthesis system, what would you go for?

Marwin H. S. Segler answered: We recommend using the algorithms that are implemented in the common code-bases, *e.g.*, Syntheseus, AiZynthFinder, ASKCOS or your hypergraph search paper.[1] Ideally these should be trained on well-curated datasets, as public datasets require extensive curation before use.

1 P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and Teodoro Laino, *Chem. Sci.*, 2020, **11**, 3316–3325.

Adarsh V. Kalikadien addressed Marwin H. S. Segler and Philippe Schwaller: Most retrosynthetic approaches do not take reaction conditions and/or additives into account. However, these usually play a large role in experimental chemistry. Are there ways to integrate this into the approach or will it remain purely based on chemical structures?

Marwin H. S. Segler replied: Reaction conditions are already handled in several works, *e.g.*, consider Coley (2019)[1] and Schwaller (2020).[2]

1 C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, *Science*, 2019, **365**, eaax1566.
2 P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and Teodoro Laino, *Chem. Sci.*, 2020, **11**, 3316–3325.

Philippe Schwaller responded: Often, many interchangeable conditions would lead to the same results. Few retrosynthesis approaches predicted reactants and reagents/conditions simultaneously, but they are much harder to benchmark. Other approaches predict likely conditions as a separate step.

Alán Aspuru-Guzik said: Marwin, congratulations on releasing this benchmarking tool! It will significantly advance the field. I think a thing that Microsoft could do to help the community is to host this tool in Microsoft Azure for everybody to use for free. It could be hosted with all the methods being benchmarked and provided as software-as-a-service. The idea of giving things away first to attract future customers is a great success story for IBM in their approach to giving access to their quantum computers. I think Microsoft can learn from this success and follow a similar generous strategy in this case. What do you think about this?

Marwin H. S. Segler answered: We agree, it is a fantastic idea that we will discuss further in-house to see if it is feasible.

Philippe Schwaller commented: One way to improve retrosynthetic tools is to have better reaction feasibility models. How would you build these models?

Marwin H. S. Segler replied: The community has only recently started to tackle this problem. One direction is binary feasibility models that predict whether a reaction will likely work or not, which have for example been used by Coley, Bjerrum, and us. Another alternative is the inclusion of forward reaction models, for example the Molecular Transformer as used by Schwaller and colleagues, or using even three models (a retrosynthesis model to suggest the retrosynthetic disconnection, a condition prediction model, and a condition-dependent forward model, as used by, *e.g.*, Coley and Reymond. We would encourage the community to further look into this challenge.

Philippe Schwaller queried: How would you make simulations general enough? Would you build simulations for specific reaction types?

Marwin H. S. Segler responded: This is a great question. As we have argued in our 2018 Nature paper,[1] we believe a desirable end goal for ML in organic synthesis is to have a unified model that can predict a wide variety of reaction types. This is of course challenging, as the scope of chemical reactions is not fully understood, and in fact also grows (consider the very recent revision of Bredt's rule, for example). Nevertheless, when the objective is to model a specific reaction in great detail, it can be highly beneficial to build models or run simulations for a single reaction type or even a single reaction only. It depends on what problem one tries to solve.

1 M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.

Christian Kuttner enquired: What are the most critical challenges in designing universally accepted benchmarks for retrosynthesis, and how might this collaborative effort accelerate the field's progress?

Marwin H. S. Segler answered: In machine learning, and most prominently in computer vision, standardized benchmarks have led to significant progress. Apart from what we have summarized in our paper (**https://doi.org/10.1039/d4fd00093e**), the challenges in designing benchmarks are making computationally tractable benchmarks that are still chemically meaningful. Such computational benchmarks are enabling machine-learning researchers with limited chemistry expertise to come up with new algorithms with confidence, and chemists who want to apply algorithms to pick algorithms with confidence.

Christian Kuttner communicated: You highlighted that current benchmarks for synthesis planning algorithms may be imperfect and inconsistent, masking underlying issues in performance evaluations. How does Syntheseus improve the accuracy and consistency of benchmarking in retrosynthesis, and what specific gaps in existing benchmarks does it address?

Marwin H. S. Segler communicated in reply: We provide an extensive discussion on this in our paper (**https://doi.org/10.1039/d4fd00093e**).

Matthew R. Ryder communicated: Your work (**https://doi.org/10.1039/d4fd00093e**) addresses the need for better benchmarks in retrosynthesis

algorithms. How do you see these benchmarks evolving to better capture the complexities of real-world systems, including multi-step reactions or materials with dynamic or disordered states?

Marwin H. S. Segler communicated in reply: Benchmarking for small-molecule multi-step synthesis is already supported by our framework, and in our work (**https://doi.org/10.1039/d4fd00093e**) we also discuss important other works, *e.g.*, PaRoutes by Genheden/Bjerrum.[1] A challenge in multi-step synthesis is that often the criteria that chemists use to come up with routes are very hard to capture formally, or into simple computatable scores, which makes them less amenable to inclusion into a benchmarking framework. This includes, for example, selecting routes based on convenient intermediates, which is contextual information that is often lacking. Materials synthesis is usually quite different from small-molecule organic synthesis, but nevertheless very important. However, as this is not my area of expertise I would refer to the excellent work that has been started by the community in this area.

1 S. Genheden and E. Bjerrum, *Digital Discovery*, 2022, **1**, 527–539.

Erwin Lam opened a discussion of the paper by Basita Das: How can you implement synthetic procedures into your pipeline?

Basita Das responded: There are always certain rules of thumb chemists use to synthesize materials. I hope to capture some of them through my conversation with chemists. Once I know what those rules are, *e.g.* some specific temperature range for annealing, certain combinations of solvents, *etc.*, I will add them to the filters such that when we are making materials to which those rules might apply, they are suggested.

Chris J. Pickard said: Your pragmatic approach (**https://doi.org/10.1039/d4fd00120f**), based on chemical experience, for filtering the results of large-scale structure searches is very appealing. It is becoming clear that, to date, these large-scale searches are rather "shallow" in the sense that there is a large chance that more stable phases, which will displace many of the predicted structures on the convex hull, have been missed. This is a "bitter lesson" in the field of computational materials discovery. The more effort you put into the searching the compositional and structural space, the less exciting your results become. It is very important to put the work into trying to displace your favoured phases, especially if they are extraordinary in some sense.

Chris J. Pickard continued: It is worth reflecting on the nature of discovery given the topic of this discussion. The term "discovery" can be used to describe different styles of scientific progress – there is discovery, and discovery. What might be referred to as "discovery light" is something that data-driven approaches excel in. Gaps in chemical space can be explored interpolatively, and properties optimised, to discover new stable phases, possibly with excellent and useful properties. This is a key task of "materials discovery". But it feels very different to "real discovery", where chemistry and physics is revealed that is at odds with existing understanding – for example, the discovery of superconductivity by

Onnes, or quasicrystals by Shechtman. Typically, this real discovery is the preserve of experimenters. However, it is possible to discover unexpected phenomena if first principles, extrapolative, searches, based on the fundamental laws of physics, are performed. Examples from our work include the anticipation of the mixed phases in dense hydrogen, and the decomposition of ammonia into ammonium amide at megabar pressures (**https://doi.org/10.1039/d4fd00134f**).

Basita Das replied: It's a very interesting question and my thought on the usage of the word "discovery" is that depending on who you ask, the meaning of "discovery" will be different. In my world "material discovery" refers to the act of "successful synthesis" of a new material validated using different characterization techniques such as XRD/XPS, *etc.* The idea is that we "predict" new material compositions using AI/ML and then physically synthesize them from available salts. A material can indeed exist in different phases, and synthesizing each of them can be termed "discovery of a new phase". The key action associated with "discovery" in this context is the physical validation.

Nicholas David commented: I appreciate your approach to capturing the intuitions of solid-state chemists, which can be rather qualitative at times, into these quantitative features for screening likely unsynthesizable materials. How many of the compounds that were filtered out were indeed present in the ICSD or known to be experimentally synthesized? In other words, for which material systems does your filtering mechanism perform poorly?

Basita Das responded: We have only tried our filtering algorithm on perovskite inspired materials. These filters can be used for other chemical spaces too, since there is nothing in them that is just for perovskites. Actually it did not filter out any compounds which exist experimentally or have a materials project ID.

Nicholas David said: I've seen other models for synthesis that produce a "synthesizability" score from 0 (least synthesizable) to 1 (most synthesizable) for computationally predicted compounds. These probabilistic measures of synthesizability have received some criticism due to the ambiguity of a single probability measure indicating whether an experimentalist should or should not attempt to synthesize a compound. For example, what does a synthesizability score of 0.8 mean? Does it mean that I'll be successful in synthesizing the compound 80% of the time? Furthermore, these models typically rely on the distribution of previously synthesized compounds, which we know to be heavily biased towards oxides and other regions of chemical space. How does your approach differ and what influenced your decision to go with these binary filters for modeling materials synthesis? And how might your filtering scheme be better at discovering novel materials in previously unexplored regions of chemical space?

Basita Das answered: Very good question. I think the way to read those probabilities is that there is 80% certainty that you might be able to make those compounds. At least that is how I interpret them. What influenced this work: I did not start my work thinking I will develop filters. I, like many others, thought that if I predict materials with AI and predict if they are synthesizable or not with convex hull prediction, I will be able to make those materials. But I was wrong, very

wrong, as synthesizing materials is a whole different story. In this process I realized that experimental chemists have a lot of subtle rules/intuition that they use to determine if a material is going to be synthesizable or not. So, this research originated from the usual synthesizability prediction methods not working. What influenced my binary filtering scheme: That's how humans think. Chemists look at a material and they are like, "No, I don't think I can make it".

The difference from other models: it's trying to capture the human knowledge and intuition that comes from experience. It will help us weed out bad compounds from massive AI generated datasets and track down compounds that have a good chance of being able to be made. Also, unlike other algorithms, the filters try to capture different kinds of rules. So, you can get a multi-dimensional validation.And how might your filtering scheme be better at discovering novel materials in previously unexplored regions of chemical space? I believe in slow growth; I think as we slowly discover more and more materials, we will get to unknown chemical spaces too. It will not happen tomorrow, but it will happen.

Andy S. Anker remarked: I appreciate your method because it is simple and effectively incorporates the domain knowledge of chemists, which stands in stark contrast to the use of large generative models. What do you see as the key advantages and disadvantages of this approach compared to generative models? Have you conducted any benchmarking to evaluate the performance between the two?

Basita Das replied: The first advantage is that it is cheap and fast. You need thousands of dollars to make a generative design algorithm work to produce millions of compounds, most of which cannot be made or are unsynthesizable. And then you can use my model to filter out those unsynthesizable compounds from the millions of compounds on a 500 USD laptop. Bigger and faster is not always better, and there is a lot of value in the knowledge earned through experience.

We did a 5 fold validation test using the MP database.

Janine George asked: When extending your approach to a wider compositional space or by using additional filters, what challenges are you anticipating?

Basita Das responded: So, the filters are not going to be 100% transferable between different classes of materials. There are always chemical space specific rules. I don't see it as a challenge, but I think very interesting inferences are going to come out of it. We will probably see a trend or similar rules coming out of different chemical spaces.

I think the major challenge will be to find rules for chemical spaces that are not well studied. Another challenge will be when an intuitive chemical rule is in opposition to a computationally predicted number.

Alán Aspuru-Guzik commented: This is excellent work! I think incorporating the intuition of chemists and materials scientists in the screening procedures is a must. Have you guys considered building a toolkit that is readily extendable to more property screens by other groups employing an API and a pip-installable friendly package? See the work on Syntheseus in this session by Marwin Segler

as an example (https://doi.org/10.1039/d4fd00093e). Even better, can a web service be developed? For example, in the MOF world, Tom Woo has a great filtration system for making sure the CIF files and associated database input files of MOF databases are correct.[1]

1 J. Burner, J. Luo, A. White, A. Mirmiran, O. Kwon, P. G. Boyd, S. Maley, M. Gibaldi, S. Simrod, V. Ogden and T. K. Woo, *Chem. Mater.*, 2023, **35**, 900–916, DOI: 10.1021/acs.chemmater.2c02485.

Basita Das answered: Thank you, Alan, for providing the resources from the MOF world. Yes, we already have the code on github. Yes, that is the direction we want to go in, where all sorts of groups can add their rules as filters and we can build a database of different chemical rules and intuition. This will democratize the space a lot, as the domain specific knowledge will be available to all.

Janine George remarked: For including additional filters, coordination environments and their assessments will likely be important. There are already several implementations to automatically analyze them, such as LocalEnv[1] or ChemEnv,[2] in open source tools such as pymatgen.

1 H. Pan, A. M. Ganose, M. Horton, M. Aykol, K. A. Persson, N. E. R. Zimmermann and A. Jain, *Inorg. Chem.* 2021, **60**, 1590–1603.
2 D. Waroquiers, J. George, M. Horton, S. Schenk, K. A. Persson, G.-M. Rignanese, X. Gonze and G. Hautier, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2020, **76**, 683–695.

Gillian James said: I have a question about the mixed oxidation state filter. I definitely understand that as a first run, this criterion excludes potentially expensive compounds to synthesize because they require more aggressive redox procedures; however, as you mention, we know that mixed oxidation states exist: they exist in many minerals in nature (magnetite, prussian blue), and as the Wadsley/Magnéli phases of vanadia, titania, *etc.* This filter may therefore exclude a great number of transition metal oxides and other potentially interesting electronic materials, so is this really a reasonable flag for the 'synthesizability' attribute? Additionally, how are you assigning oxidation state? My understanding of most oxidation state predictions is that we assign a guessed oxidation state based on the formula and charge neutrality, but until we actually look at the structure and the charge distribution, we don't really know what the oxidation state is, nor the nature of the bonding in the material (*i.e.*, ionic, covalent, or somewhere in between). How might you make this filter more agnostic to compounds that have a high tendency to form mixed oxidation states or compounds that lie on the spectrum of covalent/ionic bonding?

Basita Das responded: This is very good information, which I did not know. So far we have applied this filter to perovskite-inspired materials systems only where bonds are more ionic. The filters have to be tuned to the material system we are applying them to. So, when we search a chemical space with transition metals, we need to alter that flag. Even in filters, there is no one size that fits all and we have to use our domain-specific knowledge to tune the filters to fit the systems. But thank you for your question, it has a lot of information that will be useful in designing better filters.

Chris J. Pickard commented: You mentioned the errors reported on the Matbench leaderboard. This refers to the performance of universal machine-learning models compared to an assumed ground truth based on density functional theory (DFT). I find it most useful to think of the machine-learning interatomic potentials (MLIPs) as an acceleration strategy for DFT. At their best, the MLIPs will reach the accuracy of DFT, but at a much lower computational cost. When assessing convex hulls, given the level of errors in the current generation of MLIPs, it is important to focus on the DFT energies (which, of course, have their own errors). Research groups exploring the machine-learning-generated convex hulls might reach out to DFT practitioners to refine the energies, and in the case of restricted chemical spaces, perform deeper searches for competing phases.

Christian Kuttner communicated: While synthetic datasets and algorithms generate millions of candidate compounds, how can we better integrate a chemist's intuitive understanding of synthesizability into these automated systems? Is it feasible to develop software environments that combine data-driven approaches with human intuition in a more seamless and effective way?

Basita Das communicated in reply: That is exactly what we are trying to do. Filters capture chemist's intuition in a software environment which can be attached to a generative design algorithm which captures the learning from data. So the combination of filters and generative design algorithm is indeed the marriage of human intuition and data.

Jennie Martin opened a discussion of the paper by Aron Walsh: Have you attempted the use of clustering algorithms on the high-dimensional space from the compositional embeddings prior to dimensionality reduction, to investigate whether any of the same groupings and relationships found in the 2D mappings could also be uncovered in this way? If not, do you have any insight into what you might expect to see if you did?

Aron Walsh responded: Our strategy was to keep the labels hidden, hoping they naturally map onto one of the compositional embedding spaces. While this wasn't successful generally, we did see an increased tendency for clustering as the number of components increased. To answer your question, we didn't try that approach, but forming enhanced embeddings with our new labels is a good idea that I would like to try out.

Matthew Dyer said: Discretisation of chemical space into integer stoichiometries is an attractive starting point, but in reality the space of potential accessible compositions in the inorganic solid state is continuous (consider, for example, solid solutions and doped compositions). Is there any way that we can start to map and measure this continuous space?

Aron Walsh answered: I fully agree. My perspective is that integer stoichiometries effectively define a coarse grid that set the limits of compositional space. Stoichiometric deviations and the formation of more complex mixtures require a finer grid or the introduction of additional dimensions. Perhaps an efficient

approach would be adaptive higher-resolution sampling in regions of space where this is most important.

Venkat Kapil remarked: I loved the classification of materials into four categories. It would be interesting to check where they lie on the convex hull. This exercise will be helpful in putting into perspective recent work that uses convex hulls for the 'discovery of new materials'.[1]

1 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, *Nature*, 2023, **624**, 80–85, DOI: **10.1038/s41586-023-06735-9**.

Aron Walsh replied: The way that we can map such large chemical spaces is that this study (**https://doi.org/10.1039/d4fd00063c**) was restricted to chemical compositions only. To calculate accurate energetics, we should add a structure dimension, which for "missing" and "unlikely" compounds would require a large effort given the massive numbers involved. I do believe that generative AI may be beneficial here.

I also have a concern with standard convex-hull analysis. For example, many peculiar compositions such as CsO_8 may be predicted to be stable with respect to other crystals at 0 K, but not for the formation of gaseous oxygen. It is time to start building more sophisticated first-principles phase diagrams including temperature.

Tim Bechtel queried: Does the analysis change with different hyper parameters of the dimensionality reductions, or is this accounted for by looking at different algorithms? In my small experiments with these algorithms, changing the parameters often changes the mapping drastically. Does this affect the analysis of the mapping?

Aron Walsh responded: This is a valid point, especially for quantitative analysis of such maps. In this study (**https://doi.org/10.1039/d4fd00063c**), we were more concerned with the qualitative landscape. We chose to explore the impact of choice in element features and dimensionality reduction schemes, but deliberately kept the hyperparameters fixed. One extension could be to explore how the clustering of our labels is influenced, which may impact the targeted sampling of inorganic space.

Nicholas David enquired: To you, what does it mean for a discovered material to be novel? Recently, there have been some criticisms of the claims made using the GNoME dataset in that a large number of the discovered compositions adopt structures that already exist in the ICSD. How might we come up with a better measure of novelty in materials discovery? As AI begins to proliferate materials discovery, defining what qualifies a computationally predicted material as novel is becoming increasingly important.

Aron Walsh answered: Novelty is a point that has been raised in several of the papers and discussions so far. To human chemists, a trivial modification of a known compound may be thought to lack novelty. To a machine-learning model, the distance of a new compound to known compounds in a relevant

composition–structure–property feature space may be used as a more quantitative measure. But at this point in time, I am not aware of a general definition that has been accepted by the community and I certainly agree this is important to establish.

Alex M. Ganose commented: Often, technologically interesting materials have defects or compositional disorder. Can your approach be extended to capture these systems?

Aron Walsh responded: It is definitely possible, but with obstacles to overcome when moving away from simple bulk features to those that depend on processing. I can see how changes in stoichiometry could be readily treated using this approach. A more difficult case is stoichiometry-preserving defects, as found for Frenkel or Schottky type disorder, which would require the introduction of additional dimensions.

Heather J. Kulik communicated: Do you think the rules you have developed could be extended to crystals with organic building blocks, such as metal–organic frameworks and covalent organic frameworks?

Aron Walsh communicated in reply: The basic concept of electron counting is definitely transferable to other chemistries. I am aware that related concepts are sometimes integrated into validity checks for databases of measured and hypothetical porous materials. An additional complication can be compensating charges in pores, which are sometimes missing or neglected, and can change assignments of oxidation states and the determination of charge neutrality.

Ksenia R. Briling communicated: The compositions that do not pass the chemical filter are called "interesting" if they are present in the Materials Project database and "unlikely" if not. However, Materials Project is far from complete, since there is a huge number of "missing" compositions. Is there a way to find the compositions among the "unlikely" set that have potential to become "interesting", maybe using the discussed embeddings? Are the "interesting" composition truly interesting, *e.g.*, for some applications?

Aron Walsh communicated in reply: Thank you for the stimulating question. The labels we defined are quite soft, but I hope they can be helpful and further refined to more quantitative classes in the future. My belief is that "unlikely" may contain some of the most exciting (unconventional) compounds, but it will be challenging to identify them.

The initial set of chemical filters in **https://github.com/WMD-group/SMACT** were targeted towards heteropolar solids built from atoms in a given oxidation state. However, we have been working on generalising these rules for a better description of intermetallic systems and those with unusual chemical bonding. This would make some of the "interesting" compounds "standard", and some of the "unlikely" compounds "interesting". We have more work to do in this direction, which will include drawing from databases beyond the Materials Project.

Volker L. Deringer communicated: Your classification system and the statistics in Table 2 of your paper (**https://doi.org/10.1039/d4fd00063c**) are very interesting. Could these statistics allow you to make recommendations for colleagues who are developing large materials databases (such as the Materials Project used here)? For example, is it worth collecting more data for the existing "interesting" binary phases, of which there's more than the "standard" ones, to understand them better? Or should one prioritise the computational exploration of "unlikely" compounds, in a search for something genuinely new?

Aron Walsh communicated in reply: I fully agree. These points also link back to the concept of "novelty", which has been raised several times in the discussion at this meeting. There is an opportunity to develop a more canonical set of labels, drawn from multiple sources, that would help to direct discovery efforts. Rather than being unobtainable, "unlikely" may be one of the most fertile regions for novel findings.

Matthew L. Evans communicated: I appreciate your comments regarding the availability of underlying source data in the case of the GNome dataset, in particular. Related to this, I have a bit of a multi-part question and comment/advert! First, do you think we can incentivize data sharing through tools like OPTIMADE so that we have a way to really query within these giant hypothetical convex hulls and avoid marketing spiel about who has the largest database? After its publication, I took it upon myself to ingest the GNome dataset as an OPTIMADE API, which I am hosting at **https://optimade-gnome.odbx.science/v1/structures**, but this process should not rely on me! Secondly, it is clear from your paper (**https://doi.org/10.1039/d4fd00063c**) and discussions this week that the ability to generate new hypothetical materials is great, but the "market" is already flooded if you are an experimental group looking for promising compounds to synthesize, and thus we fall back to relying on interpersonal connections and collaborations to get things made (and the abundance of less than feasible compounds somewhat reduces our credibility as a field). However, even if the majority of hypothetical materials are perhaps unsynthesizable, I find it hard to believe that there is really nothing left to discover – do you see a way for us to handle this in a decentralized way? My own little toy idea is the tongue-in-cheek site at **thismaterialdoesnotexist.com** (referencing the famous **thispersondoesnotexist.com**), where you receive a random structure from the GNome dataset and are asked to give a hand-wavy likelihood of whether you think it is synthesizable. I believe this could lead to some interesting statistics (especially with curated "expert" user groups comprising, *e.g.*, inorganic chemistry students) and, when combined with synthesizability classifier models, may go some way to focusing experimental efforts in more fruitful areas (though currently this site remains a toy unless I can find interested partners).

Aron Walsh communicated in reply: It is a shame that the GNome dataset was not made accessible to the community at the time of publication. Thank you for your efforts with OPTIMADE, which my research group has also benefited from. In contrast to there being "nothing left to discover", I agree that the space for forming materials is vast. A potential drawback of human (statistical) classifiers is

that their knowledge (data) of what is possible is limited. I am often in awe of the compositions and structures that synthetic chemists can realise in the lab (*e.g.*, under high pressure or using reactive precursors), so we have to be sure to capture that complexity in the definition of what can be synthesised.

Graeme M. Day communicated: The results in Table 2 of your paper (**https://doi.org/10.1039/d4fd00063c**) show an increasing proportion of 'missing' structures – those that pass the filters, but are unknown – when moving from binary, through ternary and quaternary compounds. This leads to the conclusion that there is more potential for discovering new materials in the higher-order compounds. I wondered whether these increasing proportions of missing structures also reflect higher statistical probabilities that compounds with more components lie above the convex hull, *i.e.*, it seems more likely that there are more stable combinations of simpler compositions as the compositional complexity of the target material increases. This could be another explanation of more 'missing' compounds in your survey. Would you agree with this, and can you comment on whether this could be quantified?

Aron Walsh communicated in reply: This is a pertinent question, which I don't believe I can answer confidently yet. The labels I presented are drawn from entries in the Materials Project. There is a bias in this dataset for ordered crystal structures and there is also a lack of temperature in the available thermodynamic potentials. For example, if I mix two binary compounds AB + CD, they may form an ordered quaternary compound ABCD. However, the elements could also mix to form a disordered binary structure (A,B)(C,D) or a disordered ternary structure, such as AB(C,D). Such phase behaviour is sensitive to temperature and pressure, and entropy plays a decisive role. So I believe multi-component solids are underrepresented. But I do agree with your point and we are attempting to work in this direction.

Christian Kuttner opened a discussion of the paper by Wenhao Sun: You mentioned that text-mined synthesis datasets do not fully satisfy the "4 Vs" of data science. How can future data-mining efforts improve the quality and completeness of these datasets to better fulfill the "4 Vs" criteria and enhance predictive modeling?

Wenhao Sun answered: Thank you for the question. As discussed in our manuscript (**https://doi.org/10.1039/d4fd00112e**), I don't think that the limitations of the 4 Vs stem primarily from the text-mining or data-mining side, but rather, arise from anthropogenic biases in how the literature dataset has been produced. I think that chemists tend to choose materials systems to research in a more 'exploitative' than 'explorative' manner – meaning that they tend to study known materials rather than new materials. This is a less risky strategy if you want to complete a PhD in 5 years. This tendency has limited the chemical variety in our datasets; and variety is important to train robust machine-learning algorithms for out-of-sample prediction. I am very optimistic about the advent of robotic laboratories (which we have also led some work on; see ref. 1). I think that robotic laboratories can be used to perform larger surveys of chemical space at lower human cost, which reduces the 'risk' barrier to chemical exploration. I also think

that robotic laboratories can be used to not just synthesize new materials, but simply perform high-throughput experiments on other tedious tasks, for example, characterizing a phase diagram. It is not so easy to get funded to characterize phase diagrams anymore, even though it would be extremely valuable to do so. These are other sources of data that can help enhance predictive modeling.

1 J. Chen, S. R. Cross, L. J. Miara, J.-J. Cho, Y. Wang and W. Sun, *Nat. Synth.*, 2024, **3**, 606–614.

Yuchen Lou said: I admire the length you have gone to interrogate your dataset. If I have the opportunity to add more structures to my property predictor models, how do I choose the structures without falling into the anthropogenic biases that you have outlined?

Wenhao Sun replied: Well, I think you could start by adding phases from crystal structure predictions – which will generate numerous hypothetical structures that will not be subject to anthropogenic bias. The downside is of course that these hypothetical structures may also not contain the reasonable physics that govern crystal structure selection. One tradeoff may be to include common prototype structures (*e.g.*, rutile, perovskite, spinel, *etc.*) but in chemical spaces that have not yet been previously explored by humans. This balances the usage of structures that are common, but in chemistries that have not yet been experimentally explored (and therefore have opportunities for discovery of new chemically-driven properties).

Niamh Hickey remarked: In your paper (**https://doi.org/10.1039/d4fd00112e**), you mentioned that there are over 500 experimentally known ternary sulfides, but only 13% of them have text-mined recipes mentioned. Do you think that research like this would drive an experimental chemist to explore the synthesis of new compounds and to publish new recipes for these compounds?

Wenhao Sun responded: So, actually I think these recipes exist, it's just that a lot of them were published prior to the year 2000, so we don't have them in our text-mined dataset. That being said, I do think that collecting these older recipes and scrutinizing them would be very valuable. I will tell you my perspective that is not published, but is based on my experience. There are substantial differences between the synthesis of oxides and non-oxide chalcogenides (sulfides, selenides, tellurides). Many of the principles that we have made for predictive synthesis do not work very well in the sulfides. If I had to conjecture, I think it is because oxides decompose at high temperature usually into the gas phase (volatilizing oxygen), whereas sulfides often form liquid sulfur intermediates (such as the $FeS_2 \rightarrow FeS + S(liq)$ peritectic decomposition). I think this changes a lot of the phenomenology of how sulfides form, and we do not currently really have strong predictive methods to navigate this yet.

Erwin Lam asked: Could you come up with some guidelines/templates for how researchers should write their procedures to help with text mining?

Wenhao Sun answered: I think ref. 1 gives an excellent set of guidelines and suggestions for how to write a synthesis paragraph in preparation for text-mining.

That being said, this paper was written before widespread adoption of large language models (LLMs), and I think we might not be quite as constrained about how we write that paragraph anymore; LLMs might be quite flexible at picking up the important aspects of synthesis from natural language.

1 E. Kim, K. Huang, O. Kononova, G. Ceder and E. Olivetti, *Matter*, 2019, **1**, 8–12.

Venkat Kapil commented: You have raised valid concerns about the appropriateness of the currently used descriptors for accurately predicting properties. However, it seems that properties like reaction temperature, especially if you aim to predict them within ± 10 K, are correlated with the molecular structure. Given this, don't you think it would be more effective to use atomic and molecular representations rather than those based on cheminformatics? One could even argue that it is overly optimistic to expect predictions of reaction temperatures within ± 10 K using cheminformatics descriptors only.

Wenhao Sun replied: So, in my experience, I think that atomistic structure is an important consideration only in certain contexts. I think that when thermodynamic driving forces are large (*e.g.*, when $\Delta G_{\text{reaction}}$ is large), nucleation and growth processes will dominate the kinetics in such a way that the initial structure of the precursors and the final structure of the product do not have to necessarily be correlated. On the other hand, when thermodynamic driving forces are small, such that nucleation cannot occur, then topotactic or epitaxial transformations that involve structural considerations become important. There are then two tasks: one is to design solid-state reactions with larger driving forces, for example, by using higher energy precursors or by preferencing higher-energy kinetic intermediates; for example, as we did in ref. 1 and 2. The second task may be to consider structural effects and diffusion rates for reactions with low driving forces, such as *chimie douce* reactions. Further discussion can be found in ref. 3.

1 A. Miura, H. Ito, C. J. Bartel, W. Sun, N. C. Rosero-Navarro, K. Tadanaga, H. Nakata, K. Maeda and G. Ceder, *Mater. Horiz.*, 2020, **7**, 1310–1316.
2 J. Chen, S. R. Cross, L. J. Miara, J.-J. Cho, Y. Wang and W. Sun, *Nat. Synth.*, 2024, **3**, 606–614.
3 N. J. Szymanski, Y.-W. Byeon, Y. Sun, Y. Zeng, J. Bai, M. Kunz, D.-M. Kim, B. A Helms, C. J. Bartel, H. Kim and G. Ceder, *Sci. Adv.*, 2024, **10**(27), eadp3309.

Chris J. Pickard said: You bring up the difficulty of enforcing new standards, which has been seen to lead to the $N + 1$ problem of proliferating standards, as teams reach for just the standard that the community will rush to adopt. Maybe, with the rise of LLMs, we can take a different approach. If results are clearly, explicitly, reported, then translation between them should be straightforward, and potentially automated.

Wenhao Sun responded: Oftentimes people have put in a lot of time into planning out the 'perfect' ontology or standard for a certain data storage type. However, I think this is not a very productive use of time. I remember during the early days of the Materials Project, there were a lot of people around the world trying to enforce a new 'CIF-like' standard for computed data. But these efforts were largely ignored at the Materials Project. The philosophy that my advisor (Gerbrand Ceder) had was simply: do the best science, and whatever standard you

used to do the science will be adopted. I think this is true – because often what you need in a standard is not known *a priori*, it is during the process of doing next-level science that you will improve and update your standard to provide the most value. I guess there are standards like OPTIMADE now, but still I think the point holds – just do the best and most interesting science, and whatever standard you use to do it will be adopted by the community. With regards to this specific question, I agree that with LLMs, the urgency to make precise standards is diminished. It does not take that long anymore to use LLMs to process your data into a desired JSON format. Of course, if people do the hard work of producing a standard file or database format, this is valuable. But given that time is a limited resource, I think it might be more worthwhile to spend the time doing science rather than doing data management.

Yuchen Lou remarked: You mentioned that extracting information from papers is easier now because of LLMs like ChatGPT. Can we really trust LLMs on this, due to their tendency to hallucinate?

Wenhao Sun answered: This is an interesting question. While yes of course we need to double check if the LLMs are hallucinating, my experience is that LLMs tend to hallucinate less when retrieving text/ideas than when they are generating text/ideas. It is when they are asked to write essays with referenced sources that you see the most hallucinations. But if they are asked to interpret something that is already written, they tend to not hallucinate as much. In the language of Bloom's taxonomy, I think LLMs tend to hallucinate less when they do retrieval/recall, than when they do creation. But it is still true that results need to be carefully scrutinized by a human expert before high-throughput deployment of LLMs.

Sara Tanovic commented: In materials and organic chemistry we are seeing more open repositories where experimentalists can deposit their data in a structured way, such as the Materials Project and Open Reaction Database. Is there a future in which it could be a requirement for researchers to deposit their data when they publish new results, and would this be helpful to the development of synthetic planning tools?

Wenhao Sun replied: I think it would of course be valuable 'for us ML people' if experimentalists would deposit their data into these open repositories – but if experimentalists don't derive any value from it, it will just be an added burden for experimentalists, and I think they won't do a good job. To get their data, I think we should follow the model of these big tech companies like Facebook, Google, *etc.* They provide a service (social networking, internet search, *etc.*), and in the meantime they harness VAST amounts of data that users provide for them 'for free'. Of course Facebook has the most advanced facial recognition AI, they have people tagging their friends in photos, providing free labeled data for supervised machine-learning algorithms! Anyway, if we could provide a similar free service for chemists, I think they would be much more willing to deposit their data (and prepare it in a really good form) into these databases. For example, they could deposit XRD or CIF files of their new compounds and intermediate byproduct phases, and we could do some DFT calculations to see if you have serendipitously

synthesized a new high-T_c superconductor. (Probably not.) But giving that free service will help us collect the data from the users, from which we could later derive value for synthesis science. I also think that the enormous throughput afforded by robotic laboratories will also help us produce more reaction data, and we can directly code in ontologies for storing the chemical reaction data that comes out of robotic laboratories.

Victor Sabanza Gil said: The data analysis of extracted synthetic routes revealed a exploitative rather than explorative trend in synthetic chemists when performing experiments. This is likely due to the pressure to publish positive results. How could we convince lab practitioners to use algorithms like Bayesian optimization (BO), which may favour explorative approaches that may not be initially successful? (For example, some of the experiment suggestions of the algorithm may be purely explorative and lead to negative results, which is not desirable from the chemists' perspective, but necessary to learn about the optimization problem.)

Wenhao Sun responded: Two great points in your question. First, I'd like to point out that experimentalists don't necessarily 'want' to follow the exploitative trends, it is simply too risky to do exploratory materials discovery. However, when there are novel predictions being given, with a moderate chance of success, I think experimental chemists are very willing to follow those leads. On the other hand, you asked about their willingness to follow Bayesian optimization suggestions, which might explore parameter space in a way that is unlikely to yield a productive result, but will help us minimize uncertainty. I think that the academic chemists' goal is not only to discover and characterize materials, but also to tell interesting scientific stories. If there is a good story about BO being presented, I think they will still be interested. But moreover, if you could then convince someone that doing experiments in very different parameter spaces may help you more quickly converge on the best result, I think that is also a value proposition they will accept. It is like playing the NYTimes game Wordle. On the second word, you can choose either to leverage the clues from the first word and explore those clues further, or you can choose a completely different word to better explore the 'parameter space'. I think experimentalists appreciate this strategy and would be willing to do work to explore these ideas.

Andres Marulanda Bran remarked: Lately there has been the idea of using the synthesis of a material as the representation of the material itself. This synthesis procedure can be encoded as a string of text, and using vector embeddings from language models, this provides a numerical representation of the material.

These representations have been found in some reports to be useful for optimization tasks.[1–3]

What's your perspective on this idea? This has the advantage that it avoids the need to extract the material's structure from text, but directly uses the text.

1. M. C. Ramos, S. S. Michtavy, M. D. Porosoff and A. D. White, Bayesian Optimization of Catalysts With In-context Learning, *arXiv*, 2023, preprint, arXiv:2304.05341, DOI: **10.48550/arXiv.2304.05341**.
2. A. Kristiadi, F. Strieth-Kalthoff, M. Skreta, P. Poupart, A. Aspuru-Guzik and G. Pleiss, A Sober Look at LLMs for Material Discovery: Are They Actually Good for Bayesian

Optimization Over Molecules?, *arXiv*, 2024, preprint, arXiv:2402.05015, DOI: 10.48550/arXiv.2402.05015.
3. B. Ranković and P. Schwaller, BoChemian: Large Language Model Embeddings for Bayesian Optimization of Chemical Reactions, *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023, Retrieved from https://openreview.net/forum?id=A1RVn1m3J3.

Wenhao Sun answered: I am not very optimistic about this approach. Actually, I am fairly unoptimistic about this approach. Not that it can't be done technically, but rather I think it will have limited utility. Here are a few reasons, which supplement the details from our Faraday manuscript (https://doi.org/10.1039/d4fd00112e): (i) I think there is simply not enough information in a written paragraph to make causal statements about the efficacy of certain synthesis reaction parameters. For example, let's say a reaction actually completes in 21 minutes. However, a chemist runs the reaction for 12 hours, putting it in the oven at night and coming back to see it in the morning. The published paragraph says it takes 12 hours to make this material, but it actually only took 21 minutes, but we don't know that fact from the published paragraph. Someone else does the same reaction in 6 hours, someone else does the reaction for 24 hours. There is now a lot of confusion in the literature. We really want the datapoint '21 minutes', but most people don't characterize it to this precision – they get their product and they move on. You can make similar statements about temperature, *e.g.*, if a reaction is done at 500 °C but the published reactions are at >600 °C. With the published paragraph, you do not have a precise enough boundary to make statements about. (ii) Even if you text-mine the reactions perfectly, to build machine-learning models, you have to supplement the feature set with other descriptors about the material. In general, right now people use elemental features like electronegativity, atomic radius, *etc.*, or they use compound information like formation energies, *etc.* However, in a typical thermodynamic/kinetic analysis, these parameters would be fed into complicated reaction models, with hierarchical time and length scales, which by the way we actually know the forms of, but these are rarely leveraged in ML models. Our hope is that the ML model can find some representation of this complicated interplay between thermodynamics and kinetics, but again, I am not optimistic, especially because (i) there are not enough data points, (ii) the data points are not varied enough, and (iii) important parameters in the model are not reported in the paragraph, nor are they available in online databases (diffusivities, surface energies, *etc.*).

Barnabas A. Franklin communicated: In your publication (https://doi.org/10.1039/d4fd00112e), you comment on how historical bias limits diversity of data and thus hinders the ability of ML models. Do you think the lack of reporting of unsuccessful reactions also contributes to this? In what way do you think we as a community can encourage/incentivise researchers to publish 'dark reactions' or to contribute to databases such as the Open Reaction Database?

Wenhao Sun communicated in reply: Yes, I think the lack of reporting of unsuccessful reactions limits the diversity of synthesis data with which to train ML models. I don't necessarily think that we should publish 'all' dark reactions, as there is certainly trial-and-error involved in reaction optimization, and if we

published everything there might be a very low signal-to-noise ratio. But if there was a good-faith experimental effort to synthesize a material, and a diversity of experimental conditions were tried, but the result was ultimately unsuccessful, I think that should be published. Some good examples include ref. 1 and 2.

I also think that if an unusually complicated synthesis pathway was needed, the key phenomenology (the key traps and pitfalls of the synthesis) should be discussed in the paper, at least in the ESI. I don't think this is a policy change; I think it requires a cultural shift. Right now, it is common to only publish the final synthesis recipe, in a terse paragraph that only lists the reaction conditions. Journals and chemists both need to be comfortable and willing to publish longer descriptions of the synthesis. I think journals will benefit from increased reproducibility of published results. Chemists will benefit from the fact that all that hard work of trial-and-error will at least be recognized in the literature.

1 A. Narayan, A. Bhutani, S. Rubeck, J. N. Eckstein, D. P. Shoemaker and L. K. Wagner, *Phys. Rev. B*, 2016, **94**, 045105.
2 M. Acharya, S. Mack, A. Fernandez, J. Kim, H. Wang, K. Eriguchi, D. Meyers, V. Gopalan, J. Neaton and L. W. Martin, *Chem. Mater.*, 2020, **32**, 7274–7283.

Matthew R. Ryder communicated: Your work (**https://doi.org/10.1039/d4fd00112e**) reflects on the limitations of using text-mined data for synthesis predictions. Do you see opportunities for integrating machine learning with real-time experimental data to iteratively refine predictions, especially in systems where dynamic disorder or environmental effects play a role?

Wenhao Sun communicated in reply: Yes, certainly. Although I don't think we even need to resort to machine learning yet. Like we described in the manuscript (**https://doi.org/10.1039/d4fd00112e**), optimization of a solid-state synthesis recipe to a target compound with publication-quality purity can often take 30+ trial-and-error experiments. However, there are many interventional experiments we can do, such as using differential scanning calorimetry to get a sense of the temperature-dependent phase transitions during a powder-based reaction, and getting a sense of the reversibility/irreversibility of the reactions. One would have to take this data and, complemented with the experimental observations and the principles of thermodynamics and kinetics, make inferences based on the state of a reaction at a given temperature/time. These could be additional features to insert into an active-learning ML model. Another question is that of parameter space 'exploration' *vs.* 'exploitation'. If we have a model of the synthesis parameter space, the most useful experiment may not be some minor iterative perturbation on a previous experiment (*e.g.*, from 600 °C to 700 °C), but rather some drastically different experiment to minimize uncertainty in our model in certain parameter regions. I would call this the exploration paradigm. Then, after a few data points to explore the landscape, we could then iteratively refine around an 'optimization well' that we feel good about, with exploitation-based strategies. In summary, active and sequential learning during materials synthesis is a good idea. Chemists already do this, albeit perhaps not with quantitative models or guidance, and it would benefit them to bring aspects of Bayesian optimization into their processes. Coupled with an autonomous robotic laboratory set up, it may be possible to do this from end-to-end. But I think the next steps are not to build the robotic setup, but rather to do the fundamental synthesis science and

the algorithm development. Only after that point will we be in a position to really make self-driving labs.

Magdalena Lederbauer opened a general discussion: Syntheseus is an example of a well-documented project published as a python package. On another level, projects such as "ShowYourWork" make it possible to reproduce a whole publication, including all figures, text and numerical results (which takes a considerable amount of effort to streamline at first). Speaking also for the experimental, but specifically the "digital" science: what can we do as an academic community to make code (and results) more reproducible, adaptable and interoperable?

Christian Kuttner replied: This is indeed a complex problem, but an important one to address for the advancement of reproducibility in computational science. From the perspective of an editor, one immediate and practical step we can take is to ensure that authors provide the source data alongside their final articles. By source data, I mean the numerical data used for generating any plots, figures, charts, and statistical analyses[1] – this is crucial for transparency and reproducibility. The next logical step would be to encourage sharing the full raw data, but as you mentioned, this requires a well-organized structure and accessible metadata to ensure the material can be reused effectively.

One key idea that could significantly improve the reproducibility of computational research is the adoption of "Compute Capsules™".[2] These capsules would bundle the custom code, relevant datasets, and any necessary computational environments in a way that reviewers, and eventually the public, could easily access and execute on cloud platforms. This would eliminate the friction that comes with setting up code locally and provide a higher degree of transparency.

By enabling reviewers to run the code directly in the cloud without setup, we also reduce the technical barriers to reviewing computational science, making it easier for the wider academic community to engage with and validate the research. More importantly, once a paper is published, these compute capsules can be made openly available, enhancing the potential for others to adapt, reproduce, and build upon the work.

These steps – sharing source and raw data, providing structured metadata, and implementing compute capsules – are crucial for making computational science more transparent, adaptable, and interoperable across different research domains.[3]

1 **https://www.nature.com/ncomms/submit/how-to-submit**.
2 **https://codeocean.com/blog/what-is-a-compute-capsule**.
3 Simplifying data and code sharing, *Nat. Commun.*, 2024, **15**, 6380, DOI: **10.1038/s41467-024-50517-4**.

Marwin H. S. Segler answered: I see two parts. The first is a clear and concise description of the algorithm and the methods in the paper, such that the paper can (with some effort) be re-implemented by following the paper alone. Second is following best practices of open source software, using version control, releasing the environment files, and ideally all model artefacts and data. Unfortunately, especially in industry, that is not always feasible, though.

Philippe Schwaller asked: Regarding human intuition – how can it be captured? How can the rules be written? Shouldn't we be able to capture human intuition in data collected over years?

Miriam Mathea responded: Capturing human intuition is indeed an interesting and challenging task. While it may be difficult to directly capture human intuition in data collected over years, there are certain techniques and approaches that can be employed to leverage human intuition in the development of models and rules. While human intuition can be valuable, it is important to be aware of potential biases that could be introduced into the data. It is crucial to carefully select and curate the data used to train the models to avoid biases and ensure that the models are reliable and accurate. Additionally, explainable AI can be used to help understand and interpret the predictions made by the models, providing insights into how the models are making decisions and potentially revealing any biases that may have been introduced.

Christian Kuttner answered: Capturing human intuition is indeed challenging, but data collected over years can potentially approximate it by revealing patterns and trends that align with intuitive decision-making. Translating these insights into rules is a key challenge in AI and data science, but advancements in ML are bringing us closer to mimicking intuitive human processes. It's indeed an exciting area of exploration!

Marwin H. S. Segler replied: To some extent, human intuition is captured in historic data already, as (usually) no chemist would deliberately run a reaction that they think will fail (unless for control purposes). Our goal is to build AI algorithms that learn from data and the literature, and eventually become as powerful as experts. Frameworks on capturing human intuition with machine learning have recently been described, *e.g.*, in work led by Jimenez-Luna, and have been applied to synthetic accessibility by Schwaller.

Wenhao Sun addressed Alán Aspuru-Guzik: I often find the published, static figures of uniform manifold approximation and projection (UMAP) or t-distributed stochastic neighbor embedding (t-SNE) plots uninformative, especially when there are a lot of data-points. Since the *x*- and *y*-coordinates of the diagrams don't correspond to real information, when there are thousands of data points, they tend to overlap and any overarching structure becomes difficult to interpret. I feel like the real value of these dimensionality reduction techniques is the interactive aspect, where you can mouse over clusters and explore trends and relationships. However, this interactivity is not available in our modern form of scientific publishing. As data science increases, it should be increasingly important to put interactive data figures front-and-center in a publication, and not just as a supporting document. Alan, what are you perspectives on this as Editor-and-Chief of *Digital Discovery*?

By the way, if you can make interactive data figures possible at *Digital Discovery*, I will send all of my data-driven chemistry papers to your journal.

Alán Aspuru-Guzik responded: Thanks for your comment! I am very happy with how *Digital Discovery* is going as a Royal Society of Chemistry publication. The

main issue that the Editorial Board and I find is that the systems that run the RSC journals are really "static" and don't allow for any substantial innovation in the technology department. I was familiar with similar interactive plots by tools like **Authorea.com** about ten years ago! I hope that this recorded dialogue in the *Faraday Discussions* archive makes the RSC more risk-prone and likely to invest in new journal digital tools. I, for one, offer *Digital Discovery* as the main playground for this innovation!

Rob Evans communicated: Following on from Wenhao's comments about recording more than just the experiments that worked, it raises the question of "what defines a failed synthesis" – the wrong product? No product? Unexpected side products? A sticky black tar that still haunts me from my undergraduate days? These data add more insight into what went wrong and why – certainly more so than not recording it.

Christian Kuttner communicated in reply: It's so easy to focus solely on the successful experiments, but the "failures" often hold just as much value. Whether it's producing the wrong product, getting no product at all, or finding those dreaded unexpected side products (I can definitely relate to the sticky black tar!), these outcomes provide essential insights into what went wrong and why. Documenting these moments helps build a fuller picture of the process, and in many cases, these so-called failures can lead to breakthroughs or better understanding of the reactions. Recording the challenges and surprises is just as critical as the successes!

Rob Evans communicated: This comment follows on from one I made in Session 2. It will soon become mandatory to upload all data supporting publications, whether by employer, funder or publisher fiat. There's an opportunity here and now, rather like that discussed at the end of Session 2, to develop the appropriate tools to better visualise data and better support our publications. Different journal types/formats and an expanded range options for additional online support or links to visualisation tools are two that spring to mind.

Christian Kuttner communicated in reply: I completely agree that the growing requirement to upload supporting data offers a significant opportunity to enhance how we visualize and present research findings. Developing tools and expanding journal formats to accommodate these needs will be crucial for improving accessibility and impact. I'm particularly interested in exploring how we can integrate better visualization options to support this evolution in publishing.

Branko Ruscic communicated: We have posited the possibility of somehow introducing human intuition in ML in order to compensate for the intrinsic bias of ML models trained *via* published data, which emphasize successes but do not necessarily explicitly mention failures. However, it is worthwhile to point out here that human intuition is – by evolution – also biased. Humans have the propensity of making a Type I error, rather than a Type II error. In plain English, a Type I error is a false positive, and Type II is a false negative. The evolutionary pressure rewarding false positives and castigating false negatives occurred each time

a caveman heard a suspicious sound and ran up the tree although it was not necessary, but survived, as opposed to the caveman that erroneously assumed that the sound was nothing and ended up as somebody's lunch.

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 (**https://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. There are no other conflicts to declare.