Faraday Discussions
Cite this: *Faraday Discuss.*, 2025, **256**, 520

DISCUSSIONS

View Article Online
View Journal | View Issue

Discovering trends in big data: general discussion

Ricardo Valencia Albornoz, Dmytro Antypov, Gerd Blanke, Itamar Borges Jr., Andres Marulanda Bran, Joshua Cheung, Christopher M. Collins, Nicholas David, Graeme M. Day, Volker L. Deringer, Claudia Draxl, Annabel Eardley-Brunt, Matthew L. Evans, Ian Fairlamb, Kate Fieseler, Barnabas A. Franklin, Janine George, Joanna Grundy, Jay Johal, Adarsh V. Kalikadien, Venkat Kapil, Lyubomir Kotopanov, Vishank Kumar, Christian Kuttner, Magdalena Lederbauer, Andrea Carolina Ojeda-Porras, Jiayun Pang, Michael Parkes, Miles Pemberton, Branko Ruscic, Matthew R. Ryder, Ken Sakaushi, Gabriele Saleh, Brett M. Savoie, Philippe Schwaller, Bastian Bjerkem Skjelstad, Wenhao Sun, Takuya Taniguchi, Christopher R. Taylor, Steven Torrisi, Shubham Vishnoi, Aron Walsh and Ruiqi Wu

DOI: 10.1039/d4fd90063d

Adarsh V. Kalikadien opened a general discussion of the paper by Jiayun Pang: I was wondering about your broader view on the interpretability of ML models. In your paper you say: "FlanT5 seems slightly better at identifying key reagents involved in the reactions compared to ByT5" (**https://doi.org/10.1039/d4fd00104d**). To what extent is human bias involved in these interpretations? And should we actually aim at interpretable models or should model performance be our main focus?

Jiayun Pang responded: We simply examined the SHapley Additive exPlanations (SHAP) heatmap to identify tokens with strong colours and determine which part of the reaction the tokens originate from. This allows us to infer the possible contribution of reagents to the predicted product. This approach may lead to confirmation bias, where we see what we expect to see. Additionally, for the same reaction sequence, the tokens in ByT5 are more "fragmented" in a chemical sense and longer than those in FlanT5. Therefore some SHAP values may need to be summed up to directly compare with FlanT5, ensuring that we evaluate the same reaction fragment. Regarding the question of model performance *vs.* interpretability, I believe both are equally important. A well performed model is the first step and interpretability follows to help us understand how the prediction is

made. The interpretability would be particularly useful for tasks such as reaction condition optimisation and catalyst design.

Adarsh V. Kalikadien continued: As a follow up to my previous question, the current analysis of the models' interpretability was focused on relatively simple examples. Have you tried more complex examples? If yes, how interpretable was the model in those cases?

Jiayun Pang answered: Yes, we have provided two examples of more complex reactions in the supplementary information with our paper. They are not straightforward to interpret. We are working to analyse more reactions and looking at ways to systematically analyse the SHAP plot.

Andres Marulanda Bran asked: One of the key ideas behind the FlanT5 model is that it can be trained for multiple tasks at the same time, and the model is able to transfer knowledge between tasks and thus improve performance on individual tasks.

Is this the case in chemistry? Does training a model for doing reaction prediction, retrosynthesis, and condition prediction, all at the same time, improve performance in any of the tasks?

Jiayun Pang responded: It seems that the multi-task approach could offer a similar benefit to chemistry tasks, however the exact extent of improvement is not conclusive, as there have only been a few multi-task models developed so far for reaction prediction. In our work and that of T5Chem, we observe a 5–7% improvement in the accuracy of forward reaction prediction. A more substantial improvement was noted in the accuracy of retrosynthesis in our multi-task models, but this is currently being tested more robustly in our on-going work.

Nicholas David commented: There are three main types of transformer-based language models: (I) encoder models, (II) decoder models, and (III) encoder–decoder models. In your work, you decided to investigate encoder–decoder models (specifically FlanT5 and ByT5) for organic reaction prediction. What influenced your decision to investigate these models instead of models of type I and II? Would we expect to see any significant gain in model performance from type III models *a priori*?

Jiayun Pang responded: We started with encoder–decoder models because several previous studies have used this framework, which provides valuable benchmarks. Encode-only models (such as BERT) can be used for tasks such as prediction of reaction yield and reaction classification, but are not suitable for generating new sequences, such as generating the SMILES of possible products. Our approach should also work with decoder-only models too – any decoder model that can be fine-tuned can be specialised for similar tasks, given enough data. Whether encoder–decoder models offer a gain in accuracy compared to, for example decoder-only models remains to be tested.

Joshua Cheung asked: Do you think your model can be used to predict reaction intermediates?

Is it possible for your model to be expanded to predict multi-step synthesis pathways and reactions?

How does your model handle reactions that occur in equilibrium conditions?

Jiayun Pang answered: All the predictions we reported in the paper are single step predictions. While the models can be used to predict multi-step retrosynthesis, further fine-tuning is required to achieve decent accuracy. Our models cannot directly handle reactions that occur in equilibrium conditions as thermodynamics parameters is not included in the input reaction sequence. However our models may be able to predict major and minor products through a simple expansion of algorithms given enough data in training.

Ken Sakaushi remarked: Training data – how does content affect the outcome in your model? More data or more sophisticated model – which is more important?

How could the content of the training data affect the outcome in your model?

Moreover if we want to improve your system which would be more important – a larger amount of data or more sophisticated model?

Jiayun Pang replied: The USPTO datasets we use cover a wide range of reactions (*e.g.* 500 classes of reactions in USPTO_500_MT), therefore providing a rich representation of general organic reactions. Our data efficiency experiment (Fig. 6 in our article) showed that reducing the number of reactions significantly impact model performance.

Regarding your question on model *vs.* data, within a similar encoder–decoder framework as all the models we have tested (*i.e.* T5-style of models), the larger 'Base' variants of pre-trained models offer slightly higher performance than the 'Small' variants. However, the accuracy gap between Small and Base can be less than 1% (Table 1 in our article). Our data efficiency experiment demonstrated that the data size is still a key factor determining the 'performance magnitude' (Fig. 6 in our article). Based on our work, data size is definitely more important.

Magdalena Lederbauer commented: While "classical" machine learning often employs systematic methods like a grid search for model design and hyperparameter tuning, more complex deep learning models (*e.g.*, transformers) typically rely more on intuition and experience (or trial-and-error) due to computational constraints. This raises some questions about the approach to deep learning model design for chemistry:

– Are there, or can we develop, systematic methodologies for designing large deep learning models in chemical applications? How did you approach this?

– To what extent do we need "domain experts" in deep learning architecture design for chemistry, and what are the implications for model performance and scalability? To give a concrete example: how many layers or dimensions a model requires to perform optimally – or if these hyperparameters make a large impact.

I am curious about "best practices" in this particular field to advance the development of more effective/efficient deep learning models for the chemical sciences.

Jiayun Pang responded: There are many on-going efforts to leverage deep learning models for applications in chemistry, as we have seen in the conference. Personally, I am interested in how to make effective use of pre-trained large language models (LLMs), which are evolving into more powerful foundation models. The GPU resources and expertise required to train foundation models are beyond the reach of most researchers in chemistry. However, as our research demonstrates, finetuning these powerful pre-trained models is less computationally intensive and should be feasible for anyone with a consumer-grade GPU. In my view, most researchers will utilise deep-learning models in the near future through prompting and finetuning – customising a pre-trained foundation model to take advantage of its broad capabilities and specialising it on our own small dataset for a specific research project and area.

Regarding the second question, there have been many efforts to make deep learning more accessible. Platforms like Hugging Face have lowered the barrier for non-experts to pre-train and finetune deep learning models. In our work, the finetuning implementation has been adapted from relevant tutorials from Hugging Face. Nonetheless, despite significant improvement, I feel there is still a considerable technical barrier for a chemist to learn to train and finetune deep learning models. I am passionate about building bridges to facilitate greater collaboration between chemists and researchers from the computer science and AI communities.

Joanna Grundy asked: How is it possible that by processing all written information the LLM is capable of learning reactions? And it isn't without it? Was it that there were enough chemical reactions in the training data? Or has the model somehow learned to learn?

Jiayun Pang replied: The pre-trained LLMs have acquired an inductive bias for seq2seq tasks and therefore provide a solid foundation to finetune for out-of-domain sequences. With enough reactions (50 K to 500 K) to finetune the pre-trained LLMs, the models can become "chemistry-domain compatible". In my view, these models have not reached the level of "learning to learn", although there are indications that the more "powerful" the LLMs are, the better the finetuned model performs. For example, reaction-sequence finetuned FlanT5 is generally better than T5 in accuracy.

Brett M. Savoie queried: Do you think that the text-to-text based models face any long-term limitations compared to, say, graph-to-graph models? One could imagine foundation level models for chemistry that are trained on graph-to-graph tasks that would be cumbersome to translate into text-to-text tasks. Maybe with enough data and scale it doesn't matter?

Jiayun Pang answered: The emerging multimodal foundation models will be trained on data that represent molecules across many dimensions – sequence, graph and scientific context. Therefore, they should be able to connect all these dimensions of information about a molecule. So yes, I agree with your last comment that with enough embedded chemistry knowledge, a foundation model should be able to extract all relevant information about a molecule, regardless of the format in which we represent it to the model for query.

Philippe Schwaller responded: Regarding the comparison of text-to-text *versus* graph-to-graph models, while text models have shown impressive capabilities through techniques like SMILES representation, graph-based approaches offer some advantages for chemistry. They naturally encode molecular structure, topology, and connectivity that text models must approximate through linear representations. At scale it is probably easier to train text models. Moreover, text models excel at incorporating broader chemical knowledge from literature and procedures that are more difficult to encode as graphs.

Kate Fieseler asked: Where are these models failing? What is the most common failure mode for the retrosynthesis task? For example, are they outputting the wrong reaction transformations? If they can get the correct reaction transformation, are they outputting reactant SMILES that can be sanitised?

Jiayun Pang answered: The models perform well for forward reaction prediction, moderately on retrosynthesis and poorly on reagent prediction. We have not analysed the retrosynthesis result in detail, so cannot comment on the most common type of failure. In general, less than 10% of the SMILES output are invalid, indicating that the majority of the predictions are reasonable molecules. In the limited forward reaction predictions on C–H activation we have analysed so far, most of the model's incorrect predictions exhibit chemical common sense and are close to the task of prediction, such as having a different position of activation or a slightly different functional group.

Philippe Schwaller responded: For retrosynthesis failure modes, models frequently struggle with complex transformations and regioselectivity. When predicting reactions, they may suggest chemically valid but practically unfeasible routes. Problems often arise with stereochemistry and protecting group handling. Invalid SMILES are not the major problem as they can easily be filtered out.

Philippe Schwaller commented: There are different possibilities that could be correct in retrosynthesis even if you specify reactant class; still you get benchmark accuracies that are higher than 50% – how can you explain this?

Jiayun Pang replied: Our single task models achieved retrosynthesis accuracy around 40% which is consistent with previous work. Higher retrosynthesis accuracy was observed in our multi-task models. This improvement could come from two possible factors – the benefits of multi-task models and the dataset used, both of which require further validation. The multi-task models reported in the manuscript were trained and tested on the USPTO_500_MT dataset, which is relatively new (2022) and not widely benchmarked yet. We are currently curating our own datasets to further train and test the multi-task models.

Lyubomir Kotopanov said: I am interested in the C–H functionalisation fine tuning of the machine learning model. How large is the data set used for fine tuning? What is the diversity in terms of reaction classes within the data set? Finally, what is the data source? Are the reactions taken from the USPTO in this case as well?

Jiayun Pang answered: We used two C–H datasets, and they are not from the USPTO. The first contains approximately 600 reactions and was reported in a published journal paper.[1] This dataset contains only the C–H borylation reaction. We manually curated the second dataset ourselves, comprising around 200 C–H reactions from the literature. This dataset contains photocatalytic C–H functionalization reactions.

1 R. Kotlyarov, K. Papachristos, G. P. F. Wood and J. M. Goodman, *J. Chem. Inf. Model.*, 2024, **64**, 4286–4297.

Andrea Carolina Ojeda-Porras asked: Organic reactions usually can have unexpected products, that can be majority (undesired products) or minority (secondary products). Have you considered these unexpected results and how those affect the model that you are using?

Jiayun Pang responded: With the general organic reactions in the USPTO datasets, we assume it is usually the major product that has been recorded in most of the reactions. In our approach, the top 1 prediction should correspond to the most likely product, therefore the major product. It is likely that the top few hits may include side products, but this assumption cannot be easily verified due to the lack of reactions with both major and minor products both recorded for benchmarking. For certain type of reaction dataset where yields are recorded for at least a major and a minor product, we should be able to predict them through a simple expansion of current algorithms.

Matthew R. Ryder communicated: Your models perform well in chemistry tasks despite being pre-trained on language data. How could these models be expanded to handle reactions involving dynamic or disordered systems, such as frameworks or polymers, where predictions may need to account for non-covalent interactions?

Jiayun Pang communicated in reply: Our approach may work for other types of seq2seq tasks in chemistry, providing there is sufficient data (in sequence format) to finetune the pre-trained language models and that the data captures the general rules and knowledge of the application domain. However, it may be challenging to adapt our approach to polymers, as the SMILES sequence representation of the monomer captures only limited information about the polymer and does not account for factors, such as the arrangement and non-covalent interactions of monomer residues that lead to the polymer's configuration. Polymers resemble proteins and other biomacromolecules more than organic reactions. It may help to look into how proteins are modelled with deep learning models.

Shubham Vishnoi opened a general discussion of the paper by Christopher R. Taylor: Could you explain the rationale behind selecting crystal structures with Z prime ($Z' \leq 1$) in your study **https://doi.org/10.1039/d4fd00105b**), and how does utilizing space group symmetry to reduce search space dimensions improve the effectiveness of your methods?

Christopher R. Taylor answered: Our decision to select only crystals with $Z' \leq 1$ was a pragmatic one based largely on cost. Considering only crystal structures of molecules containing any of {C,H,N,O,F}, the proportion of structures in the Cambridge Structural Database that are $Z' > 1$ is around 11%. (Of course, if one considers only molecules with no rotatable bonds as in this work, then the proportion that are $Z' > 1$ decreases even more, as species in the unit cell are more likely to be symmetry-related since their conformations are very likely to be the same.) Meanwhile, the cost of doing CSP in $Z' > 1$ systems is increased considerably due to the additional degrees of freedom; even when assuming rigid molecules only, each additional symmetry-inequivalent species adds (in general) 6 degrees of freedom (x,y,z, and the rotational angles) that must be sampled in our structure search – for the same reason, we did not consider co-crystals in this work.

While $Z' > 1$ crystals, and co-crystals, can be readily (and are often) treated with CSP, in this work our focus was considering as large a set of molecular species as possible (within our flexibility and elemental constraints), and $Z' > 1$ systems are a comparatively small fraction of observed crystal structures. It would, however, be an obvious and interesting extension of this work to explore $Z' > 1$ systems (and, for instance, to see if CSP landscapes match the observations of experiment in preferring $Z \leq 1$ in general).

Regarding exploiting space group symmetry, this is an extremely valuable component of our approach with considerable benefits. By both generating and minimising (in the first instance) structures subject to space group symmetry constraints, we can significantly reduce the cost of the entire process. As you say, in the general sense this helps to reduce the dimensionality of the search space, with corresponding benefits in efficiency. In particular, our individual crystal structure optimisations are more efficient compared to those using a unit cell with no internal symmetry (as only symmetry-allowed optimisation steps are permitted) and we can also tune which space groups are sampled (and to what extent) to balance the exploration of possibility space with prioritising space groups that are likely to be observed experimentally. While in this work our intention was to carry out broad and reasonably complete CSP searches, when performing CSP to guide or assist experimental crystal structure determination or to predict crystal structures of systems with particular properties dependent on space group symmetry, being able to restrict the search to a known (or a set of likely) space group(s) is useful as it ensures maximal benefit for minimal cost.

Takuya Taniguchi asked: How long did it take to complete 1000 structures?

Christopher R. Taylor responded: The precise length/cost of a CSP calculation depends on many factors: the number of structures sampled, the level of theory used for optimisation, the space groups chosen to be sampled, and the size of the molecule (for such atom–atom pair potentials as were used here), to name a few. Since the molecules here were of varying size (in terms of atomic weight), some CSPs completed faster than others.

However, speaking on average across all 1000+ molecules, each CSP took approximately 10 hours on 8 nodes of 128 cores each on ARCHER2 (the UK's Tier 1 National Supercomputer). Each CSP run therefore took approximately 10 240 CPU hours, *i.e.* over 10 million CPU hours in total for the entire set of 1000+

molecules. As points of reference for comparison, a typical periodic DFT optimisation on a single crystal structure might take on the order of 1000 CPU hours, and the entirety of all submissions for the Sixth Blind Test of CSP methods[1] totalled over 40 million CPU hours.

1 A. M. Reilly, *et al.*, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 439–459.

Takuya Taniguchi said: Considering molecular flexibility – if you tried to apply your method to molecules with rotational bonds, what are the challenges?

Christopher R. Taylor answered: This is an extremely important area that is still under active research by ourselves and others. There are two main challenges to incorporating flexibility.

Firstly is the sampling of the molecular conformational degrees of freedom as well as the same crystalline ones as in the rigid case – the dimensionality of the problem rapidly grows and leads to enormous increases in computational cost of sampling the configuration space, which typically are dealt with by introducing further assumptions. (For instance, assuming that only certain molecular degrees of freedom such as specified torsional angles need to be sampled and the rest can be held rigid.) The second is in minimisation. The atom–atom pair potentials and multipole model employed herein assume fixed molecular geometries, and are not appropriate for simultaneously optimising molecular degrees of freedom (bond lengths, angles, and torsions) in conjunction with the crystalline ones. This means that, in addition to the higher dimensionality increasing the cost, the level of theory itself often needs to be adjusted compared to these rigid molecules (with corresponding increase in cost) to correctly describe the subtle balance of intra- and intermolecular forces when the molecular geometry can change in response to packing. In the past this has often been handled by methods such as density-functional tight-binding (DFTB) methods and DFT, and these are still commonly used, but increasingly there are opportunities for machine-learned total energy model potentials (such as the MACE variant trained herein) to make a substantial impact.

Venkat Kapil addressed Christopher R. Taylor: Does your MACE train on total or intermolecular energies like the FIT+DMA approach?

Christopher R. Taylor replied: Our MACE model is trained on the total DFT energies (calculated in VASP at the PBE+D3 level of theory) of CSP crystal structures with minor perturbations to some atomic coordinates (we describe them as "rattled" structures), to introduce some variation in the molecular geometries and thus obtain some description of the intramolecular contribution to the total energy. While it is unlikely to be able to describe very drastic changes in molecular geometry, our MACE model is a total energy model and does incorporate some knowledge of the internal energy of the molecule. More details on the training of the model are available in the ESI with our paper.

Venkat Kapil suggested: Have you tried training a single MACE model on multiple systems simultaneously and checking how it compares against individual models on different systems?

Christopher R. Taylor responded: The MACE model(s) we have trained here are indeed trained on multiple different systems, in the sense that the dataset for training consists of PBE+D3 total energies of a variety of crystal structures of molecules (selected based on being part of the previous neural network potential lattice energy correction training dataset discussed earlier in the work). They are not segregated to individual MACE models for different molecular species; rather, we trained an initial model (on the perturbed or "rattled" versions of structures taken from the NNP training dataset), and then a second phase model which incorporated additional training data in the form of PBE+D3 energies and forces of 5 further structures per CSP landscape (*i.e.* per molecule) that had been optimised using the first-phase model.

Our focus in this work was primarily on exploiting the size and breadth of the data to make a "first pass" at transferable machine-learned models, hence building our MACE model using structures from all the CSP landscapes. It might be interesting to explore the performance of fitting individual MACE models for specific landscapes (*i.e.* specific molecules), even if just to set something of an upper-bound on the performance of our MACE approach. However, that was beyond the scope of this work – our generally-trained MACE model was sufficient for our proof of concept.

Volker L. Deringer remarked: You pointed out that your data can be used to fit machine-learned interatomic potentials. There are similar efforts in building large datasets for inorganic materials and organic molecules – examples for the latter are the ANI-1,[1] and SPICE,[2] datasets. What is your perspective in this regard? Would you see your dataset merging, at some point, with data for isolated molecules and organic chemistry? Or are the challenges in the area of molecular crystal-structure prediction quite distinct?

1 J. Smith, *et al.*, *Sci. Data*, 2017, **4**, 170193, DOI: **10.1038/sdata.2017.193**.
2 P. Eastman, *et al.*, *Sci. Data*, 2023, **10**, 11, DOI: **10.1038/s41597-022-01882-6**.

Christopher R. Taylor answered: This is a somewhat difficult question. Ideally and in the general sense, it makes sense for datasets that are related (*e.g.* concerning the same molecular species) to merge to at least some degree, or at least to cross-reference each other. However, the simulation methods (and properties predicted) by crystal structure prediction are often very different and not directly related to those for datasets of isolated molecules. Moreover, establishing standards and reliable computed benchmarks is more difficult, and less mature, in crystal structure prediction (and the condensed phase more generally) than for isolated molecular species. It is also arguable that the concept of the "isolated molecule" has more meaning in the organic realm than the inorganic, or at least that it makes more sense for organic datasets to be differentiated based on the molecular *versus* solid state.

In the long run, it is probably preferable that there be centralised repositories of information concerning both organic molecules and their known and predicted crystal structures, not least from the point of view of consolidating useful data and making it accessible and more impactful. (Our institution for instance is heavily involved with the Physical Sciences Data Infrastructure, and consolidation and standardisation of such datasets is clearly of interest to such an initiative.)

However, in the near term (and in my opinion), to preserve the integrity of the data and support the fair attribution and comparison of different structure prediction workflows and approaches, I feel it is better to separate CSP data from that of isolated organic molecules.

Brett M. Savoie said: The throughput that you are achieving in CSP is really inspiring. Looking towards larger structures with more flexibility, will some further parallelized version of the current approach still work or will fundamentally new strategies need to be developed?

Christopher R. Taylor replied: This is currently still heavily dependent on the degree of molecular flexibility (and to a lesser extent the size of the molecule). For molecules of broadly similar size to those in this dataset (*i.e.* tens of heavy atoms), the current approach can and has, been applied with some modifications when there are only a few (*e.g.* 1 to 4) rotatable bonds (or other intramolecular degrees of freedom). Functionally, this is mostly analogous to running multiple parallel CSPs with different molecular conformations, as suggested. One recent example is an adaptation of our workflow to a small drug-like molecule with 4 rotatable bonds.[1]

However, as molecules get larger and in particular more flexible, the "curse of dimensionality" sets in and the number of degrees of freedom makes a complete exploration of the configuration space infeasibly expensive. The application of CSP to such molecules (*e.g.* large flexible drug-like molecules or oligopeptides, for instance) requires alternative strategies, particularly in terms of how to identify crystallographically-relevant conformations and how to efficiently and reliably describe the balance and response of molecular geometries to crystal packing forces. These are active areas of research by both ourselves and other groups working in the field of CSP.

1 M. R. Ward, *et al.*, *Cryst. Growth Des.*, 2023, **23**, 7217–7230.

Christian Kuttner communicated: Your study uncovers trends such as space group preferences and provides a rationale for the spontaneous resolution of chiral molecules, using predicted datasets. How might these new insights refine our understanding of organic molecular crystals, and what practical applications in crystallography or drug design could be influenced by these findings?

Christopher R. Taylor communicated in reply: Given the scale at which we can perform crystal structure predictions for these types of systems, part of our motivation for this work was to generate a very large, general-purpose dataset of crystal structures of rigid molecules that could be exploited by the community at large. Having demonstrated their general accuracy and validity (in terms of predicting and favourably ranking the experimentally known structures), we have made this data available for the community to explore using the full breadth of relevant techniques, rather than attempting to fully exploit this data solely in-house. We are hopeful that other groups and institutions can leverage their own expertise and tools to discover new and practical insights from within this data that we have not yet considered.

Regarding our conclusions of general trends noticed across the CSP datasets, the example of space group preferences was discussed primarily to demonstrate the consistence of our generated crystal structure landscapes with observed statistics in experimental data, *i.e.* validation of our dataset rather than new understanding. However, our analysis rationalising the rarity of spontaneous chiral resolution shows one example of how CSP allows for the consideration of (as yet) unobserved outcomes to either confirm or refute experimental "rules-of-thumb". Our conclusions also provide data-driven support of the importance of developing and applying active chiral separation techniques for obtaining enantiomerically-pure solid forms, with clear relevance in drug design. We are confident that, given the breadth (and future growth) of this dataset, there are further insights here that could impact crystallography and solid-form design.

Matthew R. Ryder communicated: In the context of your large-scale CSP approach, how do you envision this dataset influencing the development of machine learning models for systems involving disorder or significant non-covalent interactions? Are there challenges in applying these insights to more dynamic systems?

Christopher R. Taylor communicated in reply: Regarding significant non-covalent interactions (chiefly considering hydrogen-bonding, but also *e.g.* halogen-bonding and π-stacking), we have considerable experience in applying our force-field and multipole energy model to hydrogen-bonded crystals and π-stacking systems, and the model generally works quite well in terms of predicting geometries and favourable energy rankings for these systems. (Indeed, the success of the distributed multipole approach for hydrogen-bonded systems is one of its significant advantages compared to point charge models.) We are therefore confident that subsequent machine learning models trained on these structures and/or energies would perform well, though our dataset would require expansion (chemically-speaking) to properly capture common hydrogen bonding groups like carboxylic acids. Halogen bonds are less well-explored by our previous approaches, in part because of a comparative shortage of experimental data to train the empirical force fields on (and therefore less-reliable CSP landscapes to train the ML models). However, an iterative approach, building on the models trained here by re-fitting them to results from (*e.g.*) DFT calculations on halogen-bonded experimental structures, could be one place to start.

Regarding disorder and dynamic systems, these are considerably more complicated. In principle the ML models trained on this data could be applied to disordered systems *via e.g.* a supercell or cluster-based approach, and may benefit from the fact that the descriptors employed are local in nature. However, the model is trained on CSP data that is fundamentally perfectly periodic, and this could limit its performance. It may be more appropriate to use the CSP landscapes computed here as physically-sensible starting points for methods in which perturbations to the structures are used to introduce disorder, similarly to the way we perturbed atomic coordinates to model flexibility. Dynamic systems present a further challenge, though our MACE potential has been trained on both energies and forces and should therefore allow for some limited description of dynamic behaviour, but is unlikely to capture *e.g.* accurate phonons as there are no second derivatives involved in training. Again, it may be that the dataset of

multipole-and-force-field minimised CSP landscapes may be the more useful element of this work for such a purpose, as it provides a large number of sensible initial configurations for calculations that can be used to train more elaborate models.

Janine George opened a general discussion of the paper by Matthew L. Evans: If another group wanted to continue the research presented in your paper (**https://doi.org/10.1039/d4fd00092g**), how easy would it be for them to do so? What barriers do you still see?

Matthew L. Evans answered: Depending on their level of motivation, quite easy, I think! We are still missing the usual things: good documentation and tutorials, as the re2fractive package is not fully generic yet. However, the combination of open source tools like *atomate2*, ModNet and jobflow-remote, and the automatic availability of our source and resulting data *via* OPTIMADE, should make our set-up quite transferable. In terms of remaining barriers, there is still a coordination issue; we are carrying on this work ourselves, but would not at all be against another group contributing to our resulting database using the same workflow settings, however coordinating this to avoid duplication of work is not trivial. Realistically, we hope to continue extending our refractive index dataset as part of ongoing work, then bootstrapping from that into a hierarchical workflow that includes nonlinear optical materials.

Janine George added: Can we use a similar approach to collaboratively generate the large amounts of data needed to train foundation models?

Matthew L. Evans answered: Potentially, yes. Dataset construction is an art in and of itself, so many of the tools for dealing with such a large "candidate pool" should already exist. In particular, incorporating the source of a given data point is crucial when dealing with such heterogeneous overlapping data, but this is something I imagine foundation models would be well-positioned to deal with. Ultimately this shifts the collaboration from centralising on a given source of structures, to a given source of definitions. Within OPTIMADE we are now facilitating the smaller groups of databases to align on definitions, for example, convex hull distance or formation energies, which we would not expect to be perfectly consistent across databases, but provides a very useful filter within a given database (*i.e.*, give me all structures that databases consider to be close to stability, according to this metric). Specifically on the re2fractive work, one can imagine several parallel campaigns for other properties (particularly those that are prohibitively expensive to compute, even for example, all of the MP) that could then be utilised by foundation models to begin directly providing property predictions.

Dmytro Antypov commented: Sixty million entries is a very large number compared to a quarter of a million of known experimental structures reported in ICSD. This means that the database is dominated by hypothetical structures that come from multiple distributions that carry different biases. What quality control measures do you put in place to decide if these structures are physically sensible? Does the metadata allow to trace back the origin of each structure?

Matthew L. Evans answered: Our first filter is whether the database reports any kind of computed stability for their entries (*i.e.*, a self-consistent convex hull for their own database). This filter immediately removes many of the contributing data sources, leaving around 4 M relevant structures. Ultimately we are running DFT workflows ourselves for our properties of interest which will automatically screen out unphysical structures, even if in cases our surrogate ML model failed to do so, but we did not run into many issues with this, as the source databases themselves are of high quality (primarily Materials Project and Alexandria). To give a 80/20 breakdown of the source of that 60 M number, around 30 M structures are from the matterverse database, which is a very large elemental substitution study using the M3GNet ML model, and another 12 M are from the NOMAD archive, which scrapes structures uploaded as part of computational workflows of all kinds and expose them for querying *via* OPTIMADE. Of course there is no guarantee that these structures are unique across databases, so one important filter is to have a measure for uniqueness to avoid duplication of expensive calculations (in our case, deduplicating by composition and space group was sufficient).

Whether the provenance metadata is available is quite database dependent. The COD, as a database of experimental crystal structures, provides the literature reference and primary source CIF of every entry they serve *via* OPTIMADE. Several other databases do the same, where perhaps a structure was taken from the ICSD and examined computationally. However, in the case of databases really generating structures *ex nihilo*, it is hard to define the provenance. This is something we are working on within the core OPTIMADE specification, to provide a light-touch characterisation for an entry (or even a whole database) that data providers can employ to suggest whether a structure is the result of for example, a refinement *vs.* a measured pattern, a computational optimisation procedure, or if it's simply just a random collection of atoms in a box.

Vishank Kumar remarked: What kind of information can you obtain from OPTIMADE API about a composition?

Matthew L. Evans responded: Ultimately this depends on the particular database. Of the current providers, most will provide you with a crystal structure and the composition defined simply by its unit cell, though we can also express disorder in our representation. An OPTIMADE API can be used to serve an arbitrary composition-property map, without providing a structure at all. As many databases provide properties resulting from *ab initio* calculations, you can find for example, band gaps, magnetic structure, bulk moduli, *etc.*, and each database will list the definitions of these 'custom'/non-core properties on their searchable info endpoints.

Vishank Kumar added: What kind of information should we capture in databases for the future to make them reusable specially for MLP applications? How do we take care of inconsistencies of VASP settings between different training data? Should you keep the whole relaxation trajectory to generate potential; or to store more information if you wish to predict charge densities in the future?

Matthew L. Evans answered: To take your questions in turn:

The utility of keeping individual relaxation steps has been demonstrated by the universal MLPs trained on the MPTraj dataset. These steps are not explicitly accessible from the MP database, but were at least kept and eventually released. Within OPTIMADE there is a drive to provide an efficient representation of trajectory data for specialist databases who wish to serve dynamics in their API. My opinion is that inconsistent settings are fine as long as they are announced clearly (and are within the desired accuracy of the initial simulation); when training on multiple data sources the provenance should be tracked and can be made available to the model in a multi-fidelity learning approach, or at the very least, modelled with a dataset-source-specific uncertainty.

While keeping every single step is not feasible, retaining uncorrelated snapshots from the trajectory is certainly useful, especially when constructing the best (*i.e.*, most data efficient, most transferable) datasets for an MLP (*e.g.*, sampling only diverse snapshots).

Steven Torrisi asked: I greatly admire the effort that it takes to collate this many data points together, and understand the challenges that come with their heterogeneous quality; as well as your agnosticism towards letting individual communities figure out their own standards. What are some creative applications that you've thought about to use all, or a large fraction, of the ~60 M data points, or any creative efforts that you've seen?

Matthew L. Evans responded: I'll start by saying that our "collection" is no more than a decentralized formatting of the individual contributing databases. Hopefully our approach also improves applications on the "homogeneous" datasets by defining rigorous standards for access and filtering (rather than just format). By going one step further and adding the federation step (*i.e.*, simply curating a list of databases that all adopt a standard), you can really begin to imagine new applications that do not rely so much on the absolute number of structures, but on the varied sources.

One example I gave in person is that of Xerus (**https://github.com/pedrobcst/Xerus**) an XRD refinement program that uses OPTIMADE to find source structures.[1] Previously, it hard-coded API queries to various individual databases which could be replaced with a single (if slightly unwieldy) OPTIMADE query. This obviously saves developer time, but I see this as a minor side effect of the main benefit, which is that just as its crucial to be able to add new entries to an existing database (rather than working from outdated snapshots), we can now plug in entirely new data sources in a decentralized way. For example, if I, as a random individual, have an idea for a crystal structure that I think is stable (by whatever means), there is now a publishing mechanism by which I can get my structure out there "instantaneously". Of course, my structure might be bad, so it is to the data consumer to decide how to use it, but for applications where there is an additional low cost filtering step (*e.g.*, for Xerus, filter by constituent elements requested by the user, then perform sanity checks on the underlying structure before trying a cheap refinement), then this can be really powerful. I see one of the main applications of ML in this space as providing near-universal low-cost filtering steps, so I am excited to see what people come up with.

Extending the work we presented in our paper, we envision parallel property campaigns that could make use of the remaining swathes of entries that were not

relevant in this study. The real power here is that most of these structures are excluded from the very first filtering step, running on the source databases themselves, rather than laboriously downloading all 60 M structures and curating them yourself. There are a few other applications listed in our recent paper,[2] and I am really hoping more will appear as we get more data sources into the fold by lowering the barrier for others to contribute (as an example specific to you Steven Torrisi, the CAMD dataset has been within OPTIMADE for the last couple of years or so on my server at **https://optimade-misc.odbx.science**, but it would be great if the institutions generating this data would buy-in to hosting it themselves!).

1 P. Baptista de Castro, *et al.*, *Adv. The. Simul.*, 2022, **5**, 2100588, DOI: **10.1002/adts.202100588**.
2 M. L. Evans, *et al.*, *Digital Discovery*, 2024, **3**, 1509, DOI: **10.1039/d4dd00039k**.

Shubham Vishnoi asked: With rapid advancements in AI and natural language processing, do you see tools like ChatGPT being integrated into OPTIMADE to simplify querying and interacting with materials data? How might this impact materials discovery and design, especially alongside machine learning?

Matthew L. Evans responded: I envision that OPTIMADE can be usefully exploited by LLMs as it has well-defined, backwards-compatible filtering grammar, so changes to the underlying database or database-specific API will not need to be part of the LLM training data. Whilst LLMs are also helpful when aggregating data from multiple disparate data sources, the queries or data representations may be subtly different in ways that the LLM cannot account for. From a user perspective, using an LLM to generate the OPTIMADE filter string based on a human-readable query can be very helpful. In this way, I can imagine LLM agents making use of OPTIMADE data sources during materials discovery campaigns to perform targeted searches of existing materials (or the lack of materials in a given space), in ways that are more effective than querying a single database alone.

Ruiqi Wu communicated: In most cases, the refractive index is dependent on wavelength. What would you expect the performance of your model to be for predicting the refractive index beyond the visible light range, such as the mid-infrared and far-infrared region?

Matthew L. Evans communicated in reply: The density-functional perturbation theory (DFPT) workflow used to compute the static dielectric tensor (from which we extract the refractive index) indeed works across all wavelengths (or non-zero band gaps). We didn't see any significant difference in performance of our model in these ranges, though model performance here is best assessed by a rank correlation, since the model is only used to prioritise future DFPT calculations. We did not make any strict delineation between different ranges in the EM spectrum, and in fact, several of our "best" materials fell in the infrared range (see Table 1 in our paper), *e.g.*, $SnTe(PbS)_4$ has a band gap of 0.56 eV, and Te_3As_2 has a band gap of 0.7 eV (modulo systematic DFT-GGA errors); as shown in Table 2 in our article, which shows the top materials "after" filtering for other concerns,

many of these infrared materials were excluded due to other criteria (toxicity, sustainability, synthesisability, *etc.*). To view the full set of such materials, you could perform the OPTIMADE query "_mcloudarchive_band_gap <1.3&sort=-_mcloudarchive_refractive_index" on our dataset hosted by Materials Cloud Archive, which will list all suggested compounds with a DFT-GGA band gap below 1.3 eV, sorted in descending order by refractive index: **https://optimade.materialscloud.org/archive/5p-vq/v1/structures?filter=_mcloudarchive_band_gap%20%3C%201.3&sort=-_mcloudarchive_refractive_index**.

Ruiqi Wu communicated: You highlight the prediction of the absolute value of the refractive index. Would you consider using the birefringence as the matrix to capture the polarization state, which describes the difference between extraordinary and ordinary refractive index?

Matthew L. Evans communicated in reply: Yes, our approach should be agnostic to the underlying surrogate model used to rank the materials for DFT validation. We would actually be very interested in applying the AnisoNet model presented by Y. Lou and A. M. Ganose at this Discussion (**https://doi.org/10.1039/d4fd00096j**), which should give us access to the full tensor. Our ongoing work involves applying a similar workflow to the one presented here to nonlinear optical properties, building on a recent dataset of second-harmonic generation (SHG) tensors from our group.[1] In this case, we will build a hierarchical pipeline that allows us to use any pre-computed information (such as refractive index) as a feature to the top-level SHG model.

1 V. Trinquet, *et al.*, *Sci. Data*, 2024, **11**, 757, DOI: **10.1038/s41597-024-03590-9**.

Christian Kuttner communicated: The OPTIMADE API allows for federated access to vast databases of both measured and hypothetical crystal structures. In what way does the standardized representation and querying format of OPTIMADE enhance collaboration across different research groups and accelerate the discovery of new high-refractive-index materials? Did compromises have to be made in the standardised format?

Matthew L. Evans communicated in reply: The main mechanism through which materials discovery can be accelerated by data sharing in this manner is the following: a group suggesting a new material does not have to screen it exhaustively for properties of interest themselves, and can instead simply publish the structures *via* any OPTIMADE API and have them picked up by autonomous systems that perform this screening. Hopefully this will encourage wider sharing and "decentralised serendipity", as groups that remain secretive will be limiting the impact of their work.

Regarding compromises in the format, I would say the general answer is "no", given that we carefully designed a way for each database/provider to extend the format and announce it in a machine-actionable way. Of course, we could not make every field or filter operation mandatory and instead just allowed each provider to announce what they support.

Matthew R. Ryder communicated: As the OPTIMADE API and federated databases grow, what role do you see for feedback loops that connect predictive machine learning models with experimental validation pipelines? Could such systems help refine predictions for complex materials and polymers?

Matthew L. Evans communicated in reply: I think this is really the crucial next step in a lot of autonomous experimentation pipelines – rather than still needing to be seeded by an idea in a given lab, it should be possible *via* OPTIMADE (and other machine-actionable APIs) to define and monitor regions of interest in chemical space, for example, if a system is set-up to synthesize nitrides, set-up a scheduled query for new nitride structures within OPTIMADE that have not yet been observed experimentally, and attempt an automated synthesis. Ideally these systems would then report their results back as an OPTIMADE API and naturally can begin networking with each other to maximise efficiency (avoid duplication) and reproducibility (incentivise duplication). In this regard there is nothing "special" about OPTIMADE (beyond the fact it already exists and has some amount of buy-in!) but I certainly envisage similar decentralized, machine-actionable APIs will be a key part of this future (and indeed, coupling this with traditional research data management in ELNs for materials chemistry is a key motivation for my current work on datalab [**https://docs.datalab-org.io**]).

Miles Pemberton opened a general discussion of the paper by Claudia Draxl: With respect to the transferability of models, you conclude that there is a need for larger and more diverse datasets to train more transferable models (**https://doi.org/10.1039/d4fd00102h**); how important is the size aspect? If we reduce the number of data points but maintain the diversity across chemical space, how might this affect the accuracy of the models?

Claudia Draxl answered: This is a good point. For the first example, Section 2 in our manuscript, the datasets are both relatively large but when we look at the histogram of the formation energies we see that many of the datapoints have similar values. We could therefore expect that a more efficient sampling of the input data space would yield similar performance. For analogous tasks in the literature, the model performance is robust to removing datapoints from heavily sampled regions,[1] so we might expect a similar result for our task.

1 J. Qi, *et al.*, *npj Comput. Mater.*, 2024, **10**, 43, DOI: **10.1038/s41524-024-01227-4**.

Gabriele Saleh asked: You observed that the formation energies from the Materials Project are, on average, lower than in AFLOW. The Materials Project database applies empirical corrections to the formation energies of some species such as oxide. Do you think these corrections could at least in part explain the lower formation energies on the Materials Project?

Claudia Draxl responded: This is an interesting approach that we would like to try in the future. It would require a systematic comparison of common materials, potentially involving machine learning. It is hard to estimate how effective this would be.

Gabriele Saleh said: Since there seems to be a systematic difference between AFLOW and Materials Project databases, do you think it's possible to apply some sort of scaling to use the data together, for example by comparing the formation energies of compounds that appear in both databases with the same structure?

Claudia Draxl answered: This is an interesting point. We cannot exclude that this could be one source of discrepancy, but are convinced that this would not be the only one. The two databases use different convergence parameters for geometry relaxation and total energy, and effective *U* values when using DFT+U. A systematic comparison would be needed to quantify the effect of all these differences on the reported properties.

Venkat Kapil remarked: Could the difference in training across datasets be a by-product of your potential not being expressive enough? I noted that your potential takes in only distances (2 body descriptors) as inputs.

Claudia Draxl replied: Since we have not done this experiment, we cannot answer this question conclusively. We do, however, think that most of the uncertainties are due to the rather small set of common materials and the differences in computational parameters. These differences appear to bias the data. We observe that the MPEU model transfers poorly to a seemingly similar dataset, suggesting that it overfits the data and learns the bias in the particular dataset. We anticipate that adding in higher body-order-terms would make the model more expressive but also more likely to overfit the data. Therefore, it would not improve the performance when transferring the models to the other respective dataset.

Wenhao Sun asked: As your demonstration case, you use a cluster expansion for the $BaSnO_3$ band gap, which is a scalar property. However, we know that band gap depends on the direct and indirect nature, and that a fitted linear trend may become unreliable if the nature of the band gap changes (if the transition VBM and CBM become different). While you could use a non-linear cluster expansion, maybe a linear cluster expansion would also work if you did the expansion with respect to the energy of various band minima, rather than the scalar band-gap value. Could you comment?

Claudia Draxl responded: The idea of using a linear cluster expansion (CE) with respect to the energy of various band minima, rather than just the fundamental band gap, is indeed interesting. However, we would like to point out that our $BaSnO_3$ data set consists of a large variety of supercells with different shape and size, with the additional complexity of chemical disorder. As a result, their Brillouin zones are not directly comparable, and identifying equivalent points across different supercells would require band unfolding techniques. While this could potentially be done, it would not be straightforward. Moreover, different vacancy concentrations and configurations affect the band gap in a different way. This could be just a change in the band extrema, but also mid-gap states. Such diversity would not be well captured by the suggested approach.

Christian Kuttner remarked: The expressivity of ML models increases with feature complexity, but this often leads to higher computational requirements. What are the most practical strategies for balancing model complexity with available computational infrastructure to train robust, scalable models for materials science applications?

Claudia Draxl replied: There are computational requirements for model training and model inference. In general, a single inference uses a fraction of the computational resources that training requires. In our experience, popular models such as Nequip[1] for instance are often retrained on datasets (sometimes the same dataset sometimes different datasets) many thousands of times, making the cumulative computational usage of the model orders of magnitude larger. As researchers who want to mitigate the negative environmental effects of our work and save resources, we should begin to incorporate the computational requirements into our optimization. Typically, only the model accuracy is used in model selection to determine the best model. Instead, we suggest considering a weighted sum or weighted product of the model's computational requirements and its accuracy (as done previously[2]) upon model selection. Another approach would be to train the model on subsets of the total data. Overall, more work in this direction is needed.

1 S. Batzner, *et al.*, *Nat. Commun.*, 2022, **13**, 2453, DOI: **10.1038/s41467-022-29939-5**.
2 D. T. Speckhard, *et al.*, *Neural Comput. & Applic.*, 2023, **35**, 12133, DOI: **10.1007/s00521-023-08345-y**.

Joanna Grundy asked: Haven't we already seen what makes data big? ChatGPT was trained on something like 45 TB, because of the sophistication of the models we are using we need huge amounts of data.

Claudia Draxl responded: We agree that for training a large model that accomplishes many different tasks, an incredible amount of data is needed. Our work has a different focus, as we don't strive for a model as complex as ChatGPT, but rather ask the question, how much data are necessary to achieve adequate performance for specific ML tasks in materials science.

Itamar Borges Jr. commented: Even with data diversity and a large amount of data in a dataset, is it possible to find new materials or molecules with unexpected properties? In other words, interesting outliers?

Claudia Draxl replied: Indeed, we have already found interesting cases when exploring large data spaces. For instance, our previous similarity search has revealed materials that have electronic properties very similar to that of a chosen reference material.[1] We also found materials that have similar characteristics, even if they do not share common features like the crystal structure or composition.[2] In neither case would chemical intuition have suggested these candidates.

1 M. Kuban *et al.*, *MRS Bulletin*, 2022, **47**, 991–999, DOI: **10.1557/s43577-022-00339-w**.
2 M. Kuban *et al.*, *Sci. Data*, 2022, 9, 646, DOI: **10.1038/s41597-022-01754-z**.

Annabel Eardley-Brunt communicated: Batch effects are a problem for big data analysis,[1] which reflect some of the issues you have studied. Can methods commonly employed in biological fields to reduce batch effects when comparing a pair of data sets, such as batch correction methods used for RNA sequencing data, be used with this type of materials data? Or is this automatically handled in your model?

1 W. W. B. Goh *et al.*, *Trends Biotechnol.*, 2022, 40, 1029–1040, DOI: **10.1016/j.tibtech.2022.02.005**.

Claudia Draxl communicated in reply: Thank you for sharing this article with us. We understand that batch effects in the biological field refer to unwanted variations in the data. This is a similar effect to what we might expect to see in materials data from slight variations in the computational parameters or experimental settings. It is an ongoing field of research to quantify the data quality. Some of the ideas presented in the article mentioned above could be used in our field as well.

Michael Parkes communicated: It is excellent to make the tools like NOMAD Camel and Oasis available, especially if it makes it easier for experimentalists to store and share their data for others. But might this lead to quantity over quality? The need for curated data has come up a few times. As was raised earlier, would it be good to discuss and agree with experimentalists on what needs to be shared and in what form? A spectrum on its own is useless. So what metadata is needed? We should all define what the minimum expected metadata should be. But what would be ideal, what experimental conditions should be shared?

Claudia Draxl communicated in reply: Metadata are the key to FAIR data and to ensure that data are reproducible. We have been working on this from early on in the NOMAD MetaInfo,[1] starting with computational materials science data, but now extending it in various directions. We closely work together with scientists from different fields in order to define as complete a set of metadata as possible for their purposes. In the context of experiments, it is also crucial to describe the sample in detail, since, composition and impurities, growth process and sample treatment, and many other factors, can change the measured property massively. We also collaborate with manufacturers of instruments to ensure that the corresponding data can be automatically extracted during measurements or crystal growth. NOMAD CAMELS (**https://nomad-lab.eu/nomad-lab/nomad-camels.html**) allow users to control many different instruments with one and the same piece of software and store the settings (metadata) and data in a uniform manner. Also Electronic Laboratory Notebooks (ELNs) incorporated in the NOMAD data infrastructure allow users to record and share experimental conditions. All of these aspects are important components of data curation.

1 DOI: **10.1038/s41597-023-02501-8**.

Janine George communicated: Will you be using the fingerprinting tools mentioned in your paper to assess the data quality of large materials databases such as NOMAD?

Claudia Draxl communicated in reply: We are using the density-of-states fingerprint already in the NOMAD Encyclopedia for finding materials that are similar to a reference material. A Jupyter notebook on this is also included in the NOMAD AI Toolkit (**https://analytics-toolkit.nomad-coe.eu/public/user-redirect/notebooks/tutorials/dos_similarity_search.ipynb**). The MADAS framework[1] allows users to explore datasets of interest. We are currently extending our efforts towards developing fingerprints and defining metrics for other materials properties. We also use machine learning for quantifying errors coming from either approximations, computational parameters, or similar.[2] Having this achieved, we plan to introduce these error estimates in NOMAD to make data from different sources more comparable, or find subsets of data that can be used together for a specific task.

1 M. Kuban, *et al.*, *Digital Discovery*, 2024, DOI: **10.1039/d4dd00258j**.
2 D. T. Speckhard, *et al.*, Extrapolation to complete basis-set limit in density-functional theory by quantile random-forest models, *arXiv*, 2023, arXiv:2303.14760, DOI: **10.48550/arXiv.2303.14760**.

Barnabas A. Franklin communicated: In your paper, you discuss infrastructure challenges, such as storage and computation. How can we adapt to the increasing demands of big data? Are cloud based services a sustainable solution or do we require an increased number of accessible super-computers? What do we need to consider in terms of financial cost and environmental impact?

Claudia Draxl communicated in reply: A recent study,[1] has shown that some LLM models require a large amount of energy to train, comparable to the yearly requirements of small towns. This article states that Google's BERT model, which is small compared to more recent LLMs, emits roughly 284 tons of carbon dioxide when trained without hyperparameter tuning on GPUs. Besides advocating for more energy from renewable resources for cloud computing, we should also consider balancing model efficiency and model accuracy. A rather common approach is model pruning, where the model size is reduced while maintaining the same level of performance.[2] An important aspect is also publishing in more detail, the trained models, to avoid retraining.

1 Emily M. Bender, *et al.*, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, 610–623, DOI: **10.1145/3442188.3445922**.
2 P. Molchanov, *et al.*, Pruning Convolutional Neural Networks for Resource Efficient Inference, *arXiv*, 2016, arXiv:1611.06440, DOI: **10.48550/arXiv.1611.06440**.

Christian Kuttner communicated: The relationship between data size and model performance is not linear, especially when accounting for quality and diversity of datasets. How can we improve the generalizability of ML models in materials science, given that larger datasets may not always translate into better predictions?

Claudia Draxl communicated in reply: One approach here would be active learning, *i.e.*, models that also estimate the uncertainty in their predictions and add new data strategically based on this uncertainty. Sampling in these regions and then re-training the model should reduce the model's uncertainty and improve its ability to generalize. Other possibilities would be to input descriptors

of the data quality into the model or use multi-fidelity ML, where heterogeneous training data are employed.

Christian Kuttner communicated: Data veracity and consistency remain a major bottleneck in building reliable machine learning models across the chemical sciences. What new methodologies or standards can be developed to ensure that high-quality datasets gathered from varied sources can be effectively integrated?

Claudia Draxl communicated in reply: Indeed, data veracity is a major problem for data interoperability. Here, data analysis techniques, like those presented in Section 3 of this manuscript, offer ways to find interoperable subsets of data. For making fundamental progress in the longer term, we need to strive, however, for training on large data sets that contain more information than, for example, a single high-throughput study. Integrating data from different sources, in turn, requires an understanding of their precision and accuracy. Benchmarking plays an important role in this context. This not only concerns the creation of benchmark datasets – experimental and computational data from all kinds of techniques – but also benchmarking methods and tools. On the computational materials science side, a very first step in comparing density-functional theory codes has been achieved.[1] But much more is needed in this direction. In the calculations underlying this paper, the best computational parameters were used, and only one method was probed. To quantify the accuracy of different methods or the numerical precision of calculations, requires a huge effort, ideally by the entire community. We are currently running high-throughput calculations in this direction, but we also encourage other people to share, for instance, their convergences studies in the NOMAD repository. Having enough data at hand, allows us to use machine learning to quantify error estimates.[2]

1 K. Lejaeghere, *et al.*, *Science*, 2016, **351**, DOI: **10.1126/science.aad3000**.
2 D. T. Speckhard, *et al.*, Extrapolation to complete basis-set limit in density-functional theory by quantile random-forest models, *arXiv*, 2023, arXiv:2303.14760, DOI: **10.48550/arXiv.2303.14760**.

Brett M. Savoie opened a general discussion of the paper by Gerd Blanke: With the extensions to non-covalent bonding for organometallics (**https://doi.org/10.1039/d4fd00145a**), is there also the possibility for InChI to describe transition states? There are commonalities in terms of describing partially bonded structures and charge states.

Gerd Blanke answered: The handling of transition states has been briefly discussed in the InChI subcommittee but we need examples (in molfile or RXN format) to better share a common understanding of what is needed for that use case. In particular, the handling of partial charges may become an issue where the current InChI algorithm may have to be extended. Please, if possible, send me any examples to my email address

Bastian Bjerkem Skjelstad said: You mentioned that haptic bonding can now be handled by InChI. This is an important concept for appropriately describing organometallic chemistry, but how about σ-complexes? In σ-complexes, the metal

center coordinates to a σ-bond – a notable example is dihydrogen complexes. Can these be represented as well? If so, are these distinguishable from the corresponding dihydride complexes?

Gerd Blanke replied: Unlike molfiles or smiles strings with their rules for bond handling, InChI only consists of connectivity and a full description of all the surrounding atoms for each atom. That leads to the introduction of the connectivity string and the hydrogen layer in each InChI. Based on that, the metal centers coordinating to a σ-bond must be transferred into connectivity information including all H atom locations. We are currently working on the implementation of the inorganic rules and have to check whether we are able to handle these bond systems.

For the time being the molfile format is the primary input source for the InChI calculation. That may be another limitation for this type of σ-bonding. Please, contact me for a further discussion as we are looking for test examples for these compound types as well.

Ian Fairlamb queried: Is it possible to use InChI for representing transition metal clusters? From an applied catalysis perspective, speciation is critically important, especially in determining the active catalyst species under working reaction conditions (an example would be cross-coupling reactions catalysed by Pd or hydrogenation reactions catalysed by Pd). Is there the opportunity to explore the metal cluster space, particularly modelling speciation and exploring potential new feasible structures that can be married with kinetic behaviour and modelling?

Gerd Blanke responded: Transition metal clusters are part of the implementation process of inorganics and organometallics we are working on at the moment. More details will come with the next release.

Jay Johal asked: With the functionality to describe organometallic complexes as a single InChi string can you extend to describe the periodic unit cells of inorganic or MOF-like crystals as a single string as well as to aid searchability or storage for large scale studies, for example? If so, can you also mark relative positions or simply the presence of guest molecules to the structural lattices to aid guest molecule studies?

Gerd Blanke replied: That area has not been addressed by the InChI up to now. We have to work on the requirements of how this can be done. This is normally done by a working group of the IUPAC-InChI subcommittee. That means as well that we need experts in that area. If you are interested to work in such a group please contact me.

Branko Ruscic commented: We are big fans of InChI – we use InChI and InChIKey as alternative chemical species identifiers in Active Thermochemical Tables (ATcT) all the time.[1] In fact, on the ATcT website one can use standard InChI for a species search (**https://atct.anl.gov/Thermochemical%20Data/version%201.202/index.php**). However, we have found that in some cases the InChI is not necessarily unique. In particular, we have encountered some problems with gas

phase ions (both cations and anions). While in many cases the standard InChI for an ion simply looks like that of the corresponding neutral species with the simple addition of the '/q + 1' or '/q − 1' layer at the end (as one would normally expect), in some cases the correct standard InChI corresponds to the InChI for the neutral species that has one less H atom (for cations) or one more H atom (for anions), with the addition of the '/p + 1' or '/p − 1' layer (with the added complication that the rather important Hill's formula layer does not quite correspond any more to the chemical composition of the target species).

1 B. Ruscic and D. H. Bross, Active Thermochemical Tables (ATcT) Thermochemical Values ver. 1.202, Argonne National Laboratory, Lemont, Ill, 2024, DOI: 10.17038/CSE/2440256.

Gerd Blanke replied: Currently, we are working on the implementations of inorganics and organometallics that touches the salt handling as well. I would like to ask you to provide us with examples as soon as possible to let us see if there are any issues with the salt handling. Please contact me with the examples.

Branko Ruscic commented: Another open question is that apparently there is currently no standard InChI for condensed phases, particularly for solids. Can we expect future developments along these lines?

Gerd Blanke replied: This topic has not found a working group up to now. The working group defines the requirements and examples for implementations like for example, the requirements for organometallics and related test data sets. In case you are interested in starting a working group for condensed phases within the IUPAC-InChI environment, please contact me.

Branko Ruscic asked: Could you give us a brief history of InChI? I remember that in the past the first two letters of the acronym were both in upper case, standing for 'IUPAC-NIST', but now these apparently stand for 'International'.

Gerd Blanke replied: Taken from a talk by Steve Heller given at the university of Aachen (RWTH) in 2022:

1970/72 NIH Mass Spectral database initiated
1973/74 NIH/EPA Mass Spectral database
1975 EPA-CAS Registry Numbers contract for MS database
1980 EPA transfers database to NBS/OSRD
1978–1983 NBS/OSRD 5 Volume + 2 supplemental volumes of mass spectra
1985/88 NIH/EPA/NIST Mass Spectral database
1998 CAS contract not renewed
1999 InChI project conceived
2000 IUPAC conference/request for NIST to assist
2005 InChI version 1 released (a decade before FAIR)
2008 First release of the hash InChIKey
2009 Release of standard InChI; it took the original algorithm with its many variable parameters and fixed them so that interoperability between databases and resources with InChIs could be achieved
2009 InChI Trust was formed, Not for profit philanthropic charity in the UK
2011 Version 1.04 released

2017 Version 1.05 released
2017 Version 1.00 of the Reaction InChI (RInChI) released
2021 Version 1.06 released
2024 Version 1.07 released.

Christian Kuttner commented: You mentioned that the InChI algorithm, which was previously a "black box", has now been documented and analyzed in detail for the v1.07 release. How does this newfound transparency of the InChI code impact the reproducibility and trustworthiness of chemical informatics research?

Gerd Blanke responded: The enhanced transparency of the InChI code is the precondition to elaborate the InChI further. For example, to work on the enhancements in inorganics we must not change the InChIs of organic compounds. Therefore, you have to be able to identify those parts in the code that allow you to normalize inorganics according to the new rules (*e.g.* bond normalization) without modifying other structure classes. So, only transparent code guarantees that you can keep the reproducibility and trustworthiness of InChI. The other aspect is that transparent code allows other programmers to contribute in the InChI development as it becomes easier for them to identify the right place for their work within the code. Out of that, like in other open-source projects, InChI is now enabled to become a community driven development.

Branko Ruscic said: We need to acknowledge the undeniable success of InChI. It started as a competition to proprietary CAS Registry Numbers, but since a year or so ago, it seems that the CAS Service is for each species also adding the corresponding InChI and InChIKey!

Gerd Blanke responded: Thanks for that. InChI has made quite a career especially since the release of InChI 1.04 in 2011. The largest databases using InChI that are known to us have more than 1 billion compounds. PubChem itself provides nearly 300 million InChIs at the substance level, and more than 120 million for compounds. InChI has become a real international standard for compound identification.

Christian Kuttner communicated: The limitations of traditional string representations for molecular inorganic compounds, particularly the disconnection of metal bonds, present a challenge for representing these structures meaningfully in chemical informatics. What specific improvements have been proposed in the new routines for handling molecular inorganic compounds, and how do these changes affect the ability to search and categorize such compounds?

Gerd Blanke communicated in reply: The inorganics and organometallics are being completely re-worked at the moment. The goal is to generally keep the bonds to metal atoms with the exception of ionic compounds. This also will be the base to identify the stereochemistry of these compound types. Last but not least we developed rule sets to uniquely identify compounds that are represented based on different bond types like for example, ferrocenes drawn with coordinative or haptic bonds. With these changes, the InChI becomes a unique identifier for

inorganics and organometallics that will allow you to search for these compound classes properly in the sense of a full identification of the compounds.

For more details, please contact me.

Wenhao Sun opened a general discussion: In Sally Price's paper,[1] she argues that the major problem in crystal structure prediction of molecular crystals is no longer the generation of crystal structures, but rather the "overprediction" of crystal structures – that is, we generate way more candidate polymorphs than are actually observed – even if they are in the energy range of observed metastable polymorphs. We actually have the same issue in inorganic materials as well. These machine-learning potentials may help us create crystal structures, but will they help us distinguish the real crystals from the 'overprediction' crystals?

1 S. L. Price, *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.*, 2013, 69, 313-328.

Graeme M. Day responded: This is a good point. Overprediction of polymorphism is a problem in organic molecular crystal structure prediction and the temperature-free approach is one reason why we predict many more structures than are observed. Some of the potential energy basins on crystal energy landscapes are quite rugged, having several local energy minima that are separated by small energy barriers that would be overcome at finite temperatures. We saw evidence for this in a study we did with Michael Shirts' group:[1] we put known crystal structures in molecular dynamics simulations and quenched the trajectories periodically by lattice energy minimisation. For some of the molecules studied, we end up in many distinct local minima, even when the molecular dynamics is performed at fairly low temperatures. All of these local energy minima correspond to the same observable crystal structure, but would be identified as unique structures when performing crystal structure prediction. There are two approaches that I am aware of that have been applied to address this problem for organic molecular crystal structure prediction. One is to perform molecular dynamics simulations on predicted crystal structures and monitor which pairs or groups of structures interconvert. Mooij's team first applied this to the simple system acetic acid[2] and more elaborate approaches using molecular dynamics have been developed by Salvalaglio and co-workers[3] In our research group, we developed a Monte Carlo approach to identify structures with low energy paths between them,[4] which reduces the number of unique structures from crystal structure prediction by grouping them into energy basins. The Monte Carlo approach is also useful for characterising the structure of the crystal energy landscape and identifying particularly deep energy basins, which correspond to kinetically stable crystal structures.[5] I think that these tools will be important for distinguishing real from 'overprediction' structures and the development of machine learning potentials will help improve the energy model on which these methods are applied.

1 E. C. Dybeck, D. P. McMahon, G. M. Day and M. R. Shirts, Exploring the multi-minima behavior of small molecule crystal polymorphs at finite temperature, *Cryst. Growth Des.*, 2019, **19**(10), 5568–5580.
2 Mooij, *et al.*, Crystal structure predictions for acetic acid, *J. Comput. Chem.*, 1998, **19**(4), 459–474.

3 N. F. Francia, L. S. Price, J. Nyman, S. L. Price and M. Salvalaglio, Systematic finite-temperature reduction of crystal energy landscapes, *Cryst. Growth Des.*, 2020, **20**(10), 6847–6862.
4 P. W. V. Butler and G. M. Day, Reducing overprediction of molecular crystal structures *via* threshold clustering, *Proc. Natl. Acad. Sci. U. S. A.*, **120**(23), e2300516120.
5 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.

Christopher R. Taylor added: This is a very good question. In Sally Price's paper, one of the hypotheses for this overprediction is that many CSP procedures use zero-temperature energy models, and that these many "false" minima are an artefact of ignoring thermal contributions. The suggestion is that accurate inclusion of thermal effects will reveal many of these zero-temperature minima to in fact be unstable at finite temperatures, with thermal motion able to overcome very low energy barriers between them and to "condense" them into the same related minimum (*i.e.* they belong to the same parent "basin" on the potential energy surface).

As provided in a comment by contributor Graeme M. Day, this has previously been studied by himself and one of his co-authors.[1] In that previous work, a Monte Carlo threshold algorithm approach was used to explore the connectivity between different local minima on the potential energy surface, and thus identify the energy barriers between minima on the CSP landscape. While it was applied to relatively simple molecules (benzene, acrylic acid, and resorcinol), it was demonstrated that the energy barriers between the many minima on the zero-temperature CSP landscape are very often less than the available thermal energy in typical experiments, and clustering structures based on their connectivities to the lowest connected minimum so determined reveals for example, for benzene, two clear "parent" minima, which correspond to the observed crystal structures.

Regarding the role of machine-learned potentials, one obvious possibility is to improve the speed and accuracy of energy evaluations in such a Monte Carlo approach. In that previous work, for the case of resorcinol, it was necessary to carry out density-functional tight-binding (DFTB) calculations to allow molecular flexibility. If transferable machine-learned potentials can improve both the accuracy of the energetics and allow some description of intramolecular flexibility at a cost lower than DFTB, this could expedite the computation of these disconnectivity graphs in general and make it more feasible to incorporate them into standard CSP workflows, reducing the overprediction of possible structures and allowing us to identify the lowest-energy "parent" structures that are likely to be observed at finite temperature.

1 P. W. V. Butler and G. M. Day, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2300516120, DOI: **10.1073/pnas.2300516120**.

Graeme M. Day said: The development and maintenance of large databases is a great thing for our field, and we're making use of these individually and, now, gathering data from multiple databases for doing our science. How do we ensure that the people who are creating the data still get the credit for their work. While the database curators are giving a great service to data science, credit needs to be fairly attributed when we use the data. Do you agree that this is a potential problem and what can we do about it?

Jiayun Pang responded: I agree that recognising data generators more effectively would benefit the field. A separate data citation alongside the more traditional publication citations could be a solution. Take organic reactions for example, the first step could be to require the deposit of reactions in a machine-readable format to a database as part of the journal submission, similar to what we do with small molecules and protein sequences and structures. In this database, all new published reactions would be searchable and have a citation index reflecting how many times they have been used to form a dataset. The database could also embed AI-driven reaction prediction models, enabling predictions based on selected reactions. This "date citation index" would give organic chemists credit for their data contribution and the benefit of interacting with AI models for reaction prediction would provide additional incentives for organic chemists to deposit their reactions.

Christopher M. Collins added: This is very much a problem. Especially as it does not offer encouragement to synthetic research groups which may have numerous compounds which have been discovered, but never published, for various reasons. It would be good, if there were encouragement for people to be able to be able to simply deposit structures within a database with a description as to how they have been made or predicted (stable or not), and still receive some sort of reward/credit.

I do wonder if such encouragement may come from having some sort of metric in services like ORCID, detailing the numbers of entries deposited within databases and/or the number of times those entries are cited in other works, in the same way that databases are starting to record the numbers of paper reviews. I of course, have no idea about the practicalities of setting something like this up!

Matthew L. Evans agreed: I definitely agree that this is a problem, relating both to conscientious scientific practices (citing your sources) and the lack of appropriate credit mechanisms. First, on the technical side, any database worth its salt should be able to provide you, automatically, with any bibliographic reference related to a given entry. This is something we baked into the OPTIMADE format (the core data types are only crystallographic structures and bibliographic references), so it is in principle possible to programmatically generate your bibliography. In cases such as our paper, where we technically "use" all 60 M structures in OPTIMADE, it is of course not feasible to cite every source publication (or even every source database), so some sensible criteria needs to be in place to for example, cite the most promising materials and the most relevant databases, to that given piece of work.

Secondly, and perhaps from a more personal viewpoint, we typically want to use credit *via* citations to estimate impact *post hoc*. I don't see any better way of "generating" impact than making your work available in open databases. In that case, I have faith that even the current flawed system would be able to eventually convert this into credit, especially if the timeline is so easily traceable *via* the data curation process. I think the same argument extends to the data curation process itself; certainly the larger initiatives (ICSD, CSD, MP and many others in our field) receive justifiably large credit for the great work they do, and I hope communities and ecosystems like OPTIMADE can extend that to smaller data providers too.

Finally, there is the question of how these things get funded in the first place, which is more tricky. OPTIMADE, for example, never received traditional grant funding beyond support from CECAM to run workshops. I think it is important to not lay out too high a bar for such a curation process, and to think about ways this work can be decentralized/federated (and thus more robust to changing funding whims) by design. There is a long payback time for such endeavours so patience is also key!

Christopher R. Taylor answered: This is a very important question – I agree this is a problem. It is not clear that the traditional publishing and citation-indexing model is well-equipped to handle for example, the citation and attribution of circa 1000 different crystal structure determinations as in our work, and such issues are only becoming more common as "big data" approaches continue to expand. Perhaps the community can come together and support the design and deployment of some mechanism by which for example, publishing platforms allow for the citation of (very large numbers of) data sources which are then correctly linked back to the original creators. As a rough example, scientific publications involving very large datasets might have a traditional "References" section for scientific references, and online formats have an additional "Data sources" section which can (and are required to) point directly to the creators of the data *via e.g.* DOI or similar. However, this is a complex problem and I agree that it requires significant consideration in the near future to ensure that data creators remain incentivised to make their data available and usable and are appropriately recognised for their contributions.

Janine George addressed Claudia Draxl and Gerd Blanke: Do you have any recommendations for the sustainable development of larger code projects in our community? Especially in terms of long-term support for such projects.

Claudia Draxl answered: Long-term funding for code development is notoriously difficult to obtain. Publishing open-source software projects offers an avenue to publicize tools in the community and potentially attract new contributors. In our experience, organizing hackathons and tutorials can also help engage the user community. Overall, we advocate teamwork over competition.

Gerd Blanke answered: The current InChI code consists of nearly 100 000 lines of C code that are sparsely documented in some parts of the code, and make the code appear as a black box for other programmers. To make it sustainable we started with 2 programmers and asked them to comment on the code they work on. This way we slowly work through the code. Meanwhile, two other (external) programmers have contributed to the code based on the documentation changes we did. Based on that our recommendations for an open-source project are:

Use GitHub or a similar provider to build an open-source repository where you store your code and work

Make sure that the code guidance is in place and transparent

The code owner has to take the decision if changes in the code are accepted

Provide sufficient automated testing for the code

Ideally, the test examples of automated unit tests cover the full (chemical) space your program works in but must be short enough to run over night, for example

For release tests, use large test sets to extensively handle the work space you are in

Provide at least 2 persons working with the code from your side to ensure that the coding and contextual knowledge is shared between multiple persons

Add extensive documentation to keep it understandable

Aron Walsh asked: Regarding the future of computational materials databases, do you think it is more beneficial to focus on collecting large datasets from many groups or on more curated datasets calculated consistently?

Claudia Draxl replied: Both approaches have their merits. Curated datasets allow for easier interpretation of the results, since they often contain less noise. However, for thinking out of the box and making major breakthroughs, we should make use of as much information as possible. For this, we need to bring together data from different sources.

Ricardo Valencia Albornoz addressed Claudia Draxl and Gerd Blanke: Just for the record, what are the main trends in the use of FAIR tools within the community? Are younger students more eager to adopt new technologies? How can the penetration of FAIR tools, such as InChI, be increased in industry, where standards are more heterogeneous or come from proprietary sources?

Gerd Blanke responded: FAIR tools seem to be primarily supported by those people that work on or are responsible for data migrations between different sources, like analytical systems or measurements of biological systems. Their work is pretty much dependent on the quality of the FAIR implementation. In terms of organizations: an organization takes more care of FAIR if its business is data related like for example, pharmaceutical companies. Accordingly, you find FAIR programs in most of the life science companies lower costs as for example, they avoid the repetition of experiments. The most interesting question in this context is whether there will be an inter-company FAIR standard. IUPAC is working on standards for the digital chemical identification where InChI will play a central role.

Claudia Draxl replied: The adoption of the FAIR principles and tools by the community is rapidly increasing. In particular, the younger generation is more familiar with electronic lab notebooks (ELNs) and eager to reuse data and tools generated by others and share theirs. For instance, NOMAD (**https://nomad-lab.eu**) has registered users from more than 1000 academic institutions worldwide. Note, however, that there are many more users, since the platform requires registration, only if uploading data. It is also important that students are familiarized with FAIR tools in their studies. Therefore, we have started very successfully to introduce ELNs in our physics lab courses.

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. There are no other conflicts to declare.

