Faraday Discussions
Cite this: *Faraday Discuss.*, 2025, **256**, 177

DISCUSSIONS

View Article Online
View Journal | View Issue

Discovering chemical structure: general discussion

Alán Aspuru-Guzik, Tim Bechtel, Varinia Bernales, Philip C. Biggin, Filippo Bigi, Itamar Borges Jr, Ksenia R. Briling, Joshua Cheung, Christopher M. Collins, Kevion K. Darmawan, Nicholas David, Graeme M. Day, Volker L. Deringer, Claudia Draxl, Matthew Dyer, Annabel Eardley-Brunt, Rob Evans, Ian Fairlamb, Barnabas A. Franklin, Janine George, Mark Goulding, Joanna Grundy, Roohollah Hafizi, Matthijs Hakkennes, Niamh Hickey, Gillian James, Veronika Juraskova, Adarsh V. Kalikadien, Venkat Kapil, Heather J. Kulik, Vishank Kumar, Christian Kuttner, Magdalena Lederbauer, Yuchen Lou, Eltjo Mante, Liam Marsh, Jennie Martin, Clelia Middleton, Tahereh Nematiaram, Charles W. P. Pare, Bianca Pasca, Chris J. Pickard, Branko Ruscic, Matthew R. Ryder, Brett M. Savoie, Wenhao Sun, Filip T. Szczypiński, Takuya Taniguchi, Steven Torrisi, Shubham Vishnoi, Aron Walsh and Shirui Wang

DOI: 10.1039/d4fd90061h

Adarsh V. Kalikadien opened the discussion of the introductory Spiers Memorial Lecture by Alán Aspuru-Guzik: You talked about the fact that both breadth and depth of research are important in addressing scientific questions. This is reflected in the range of topics that were discussed in your lecture. What motivates you to answer these questions in such varying fields ranging from catalysis to the philosophy of understanding? Are there no limitations to your interests?

Alán Aspuru-Guzik answered: Many thanks for your lovely question. I only live once and don't believe in reincarnation. I had a car accident when I was 19. This has given me an apparent urgency of purpose that joined with my curiosity, and having a dynamic and energetic research group, so that I think that *research is research* and that discipline boundaries are artificial constructs. The perspective I wrote for this *Faraday Discussion* (**https://doi.org/10.1039/d4fd00153b**) as well as a perspective I co-wrote with Markus Reiher and Roland Lindh can guide you to the general types of questions I am interested in.[1] These questions are, in general, related to the idea of accelerating scientific discovery using technology at the interface of chemistry with other fields, especially computer science and robotics.

For example, in this Faraday lecture, I didn't touch upon my group's long history of contributions to the field of quantum computing.

1 A. Aspuru-Guzik, R. Lindh and M. Reiher, *ACS Cent. Sci.*, 2018, **4**, 144–152, DOI: **10.1021/acscentsci.7b00550**.

Niamh Hickey asked: When do you think that we will have an independent artificial intelligence (AI) scientist? One that can make independent scientific decisions.

Alán Aspuru-Guzik replied: It depends on what you mean by scientific decisions. In a sense, we are already there with Bayesian optimization and hypothesis generation or testing systems at an early stage. I think you mean a full AI scientist who can create an independent research agenda. I suspect we will see this within the same decade, but as with all predictions, take it with a grain of salt. I want to end this answer with the famous quote, "The most reliable way to predict the future is to create it", which is attributed to Abraham Lincoln. We should work together as a community to make this happen as soon as possible.

Charles W. P. Pare said: Part of your talk was about the taxonomy of research subjects. In this context you introduced android scientist/AI scientist, I ask the question whether the AI scientist can be called a scientist or whether the person who asks the questions that the AI "scientist" then solves, *e.g.* through a self-constructed experiment, is the scientist?

Alán Aspuru-Guzik answered: I believe the scientist will be the one who asks the question, and it will be helped or composed by the AI scientist, at least in terms of credit. But more technically, we recently put out a preprint that can help answer this question.[1] This is my first philosophy paper, which I co-wrote with the excellent Zijian Zhang and Sara Aronowitz, a computer scientist and a philosopher, respectively. This paper discusses agents of understanding and how they can be seen as composable systems. The agent or subject could be a composite subject of a human + AI scientist, which together are more able to understand than the human alone. The paper provides a compelling argument for understanding this scenario. An example that is easy to discuss is, for example, the fact that you, plus the GPS in your phone, are better navigators, on average, than you by yourself when you are not too familiar with the environment. The paper is in preprint form.[1]

1 Z. Zhang, *et al.*, *arXiv*, 2024, preprint, arXiv:2408.08463, DOI: **10.48550/arXiv.2408.08463**.

Matthijs Hakkennes enquired: Do you believe that language models such as ChatGPT or Claude AI can truly really understand concepts, or are they just a game of statistics.

Alán Aspuru-Guzik replied: No, the current versions of ChatGPT, Claude, Llama or similar systems do not really "understand", as per the definitions from the paper with Zijian and Sara,[1] as I discussed in my previous answer. Having said this, one can imagine building systems that use these language models as

components that begin to understand given our formal definition of understanding.

1 Z. Zhang, *et al.*, *arXiv*, 2024, preprint, arXiv:2408.08463, DOI: 10.48550/arXiv.2408.08463.

Tahereh Nematiaram asked: Given your pioneering work in materials discovery, how do you envision AI's role evolving in the next decade for accelerating the discovery of sustainable materials, and what key challenges in data-driven molecular discovery/design still need to be addressed to ensure AI-driven discoveries are both efficient and environmentally responsible?

Alán Aspuru-Guzik responded: Our main challenge as a community is to build as many self-driving labs as possible openly. As we build more self-driving labs, the learning curve for them helps us make them faster, cheaper, and more reliable. The more modular we design them, the more others can replicate them. Materials characterization seems to be the bottleneck and challenge for many self-driving approaches so innovation in "proxy" approaches for measurement is always welcome.

Ian Fairlamb questioned: As an experimental synthetic chemist embracing computational methods and robotic systems in our laboratories, one of the challenges is convincing the robot developers (on software and hardware) to be interested in chemistry. How do we change that?

I am not yet convinced that human-like robots are the answer to executing experiments in a lab already designed specifically for humans to move around in. It is hard to beat (presently) the dexterity of the human hand when using a standard Schlenk line, for example. Should we be focusing on more robust automated liquid handling units and flow set-ups rather than human-like robots? A smaller scale breakthrough might be an autonomous glove (dry) box. Would we be better placed to put effort into smaller scale autonomous development of synthetic laboratories? Designing a fully autonomous synthetic chemistry laboratory from scratch would be a different discussion.

Alán Aspuru-Guzik replied: We must do *both* approaches. Firstly, most labs worldwide have equipment designed to fit humans. Secondly, most chemical reactions are also intended to be carried out in regular glassware, not high-throughput well plates or flow systems. Therefore, it would be ideal to translate all our chemistry procedures into robot-handled procedures.

I feel that economies of scale strongly favour robotic arms. Robots that replace human-like tasks are needed in all industries and are only meant to keep becoming cheaper exponentially faster. On the other hand, specialized equipment for high-throughput chemistry can cost up to millions of dollars. We can make these arms with a combination of computer vision, language models, vision models, planning algorithms, *etc.* Mobile robots and, ultimately, humanoid robots carry out the tasks we carry out in the lab every day in a human form factor.

On the other hand, companies like Opentrons have done wonders in democratizing aspects of high-throughput equipment. This equipment will be indeed helpful and accessible to us as chemists and should be part of the arsenal.

Rob Evans communicated: In your presentation, you cited Philip K. Dick's *Do Androids Dream of Electric Sheep* and talked further on AI chemists. Will it matter that the research was wholly or partially completed by an AI colleague? Do you think there will be the equivalent of a Voigt–Kampff test for AI chemists? What do you think it might look like?

Alán Aspuru-Guzik communicated in reply: Thank you for your question! Firstly, glad you found the reference to Philip K. Dick that was referred to in the presentation. To answer the first part of the question, I feel that scientific advances, no matter who or even 'what' comes up with them, are useful for the process of the field. In terms of credit, I feel that the AI systems that were developed/trained/finetuned or employed by a human are secondary to the human, at least as of the current status quo.

With regards to the Voigt–Kampff test and what it would look like, this is very related to the paper that I cited in the relevant slide.[1] In that paper, we discuss the two tests one would need to "administer" to an AI to determine if they can understand. Firstly, the AI should be able to figure out conceptual relationships and then apply them in a distinct context. Second, it should be able to explain to a human in natural language so that the human can explain it further. More details are in the paper.

1 M. Krenn, *et al.*, *Nat. Rev. Phys.*, 2022, **4**, 761–769, DOI: **10.1038/s42254-022-00518-3**.

Itamar Borges Jr communicated: Nice talk, good ideas for thinking over. The money necessary for research in AI/ML has been scaling up to the point that major billionaire companies are doing considerable methodological and application research. How do we continue to do frontier basic and applied chemistry and materials science research in the university without accessing this amount of funds or resources? Despite the significant advances in the area of AI/ML in chemistry and materials science made by research groups in the universities, including your group, are we destined, in most cases, to work on the fringes of the major problems?

Alán Aspuru-Guzik communicated in response: In my talk and associated paper (**https://doi.org/10.1039/d4fd00153b**), I address this issue. Companies and large research groups certainly can solve problems that require scaling due to the large amount of computer time or experimental resources needed. Groups with more modest resources can still make significant contributions, especially in the categories that I named Breadth and Depth in the paper, as these may not require large amounts of computer resources.

On a separate note, we have developed a review of low-cost self-driving laboratories, which are an excellent opportunity to try out worldwide.[1] They cost as little as a few hundred bucks.

1 S. Lo, *et al.*, *Digital Discovery*, 2024, **3**, 842–868, DOI: **10.1039/d3dd00223c**.

Clelia Middleton communicated: You presented an end-to-end experimental design and execution robot which, if I understand correctly, works principally based upon large language model (LLM) architectures. LLMs are notorious for

hallucinating (giving answers which initially scan as lexically reasonable but are, in fact, nonsense), and in arenas where their actions are consequential with regards to, for example, safe lab practices, this introduces a very broad and hazardous risk factor. An experienced scientist would probably be able to counter any truly egregious mistakes, but humans are not infallible, and of course the bot will be most useful where it is able to aid scientists who have not developed precise and acute domain knowledge (perhaps even eventually with students!) What methods do you suggest for risk management around hallucinations by this lab-bot during application; alternatively, how do you plan to integrate fail-safes to avoid hazardous hallucinations into the overall design architecture? This will of course continue to be a relevant challenge during the continued development of the conceptualized fully intelligent AI scientist so I would be very curious to know your assessment of the level of risk presented by hallucinations and the general philosophy of risk management on this front throughout the course of your future work.

Alán Aspuru-Guzik communicated in reply: We indeed care deeply about your question. At the moment, we employ CLAIRIFY, a tool that we developed to partially address these issues.[1] CLAIRIFY employs a verification system to make sure the robot's instructions are valid. CLAIRIFY can also be expanded to include safety considerations, which is something we plan to do. Varinia Bernales is leading a collaboration with my group on a manuscript on self-driving labs and safety. There is no DOI to share yet, but if you search our names together, you should be able to find the manuscript as a preprint as soon as we finish it in the next few weeks.

It is well thought that, at the moment, LLMs alone are not the answer. Still, the answer involves using agentic formal verification systems to check their answers against verifiers or digital twins. Our startup Axiomatic, Inc. is taking this approach to design photonics and electronics circuits.[2]

1 N. Yoshikawa, *et al.*, *Auton. Robots*, 2023, **47**, 1057–1086, DOI: **10.1007/s10514-023-10136-2**.
2 **https://www.axiomatic-ai.com**.

Filip T. Szczypiński communicated: Thank you very much for your lecture and guidance on how to identify significant research projects in the field of AI for chemistry. One of the major research themes within your group is development of algorithms for quantum computers. Developments in the quantum computing community will soon make many chemistry problems that we strive to address redundant and exact answers will be within reach for us with the knowledge of the solutions to the Schrödinger equation and the appropriate operators. What are your thoughts on the future of data-driven methods in the era of quantum computing when all our computational benchmarks and theory levels that we are striving to approximate become obsolete?

Alán Aspuru-Guzik communicated in response: That is an excellent question, Filip. Many thanks for making it! Also, I'm sorry we didn't have a chance to meet appropriately during the workshop. I enjoy the dynamics of a *Faraday Discussion*, but meeting everybody I want to is too hard.

I work on quantum computing, robotics and AI as parallel paths to accelerate molecular discovery. Each one has a *complementary* role in the puzzle, and no one of them will replace the other in its entirety, not even computational chemistry on a classical computer. There is a growing set of scientists who believe that density matrix renormalization group (DMRG) and other techniques like that will always be better than quantum computers. I am in the opposite camp. We will have to wait until we have proper error-corrected quantum computers to see who is right. Let's assume that it is worth running a quantum computer for doing *exact* quantum chemistry. I imagine that it will be more costly to do so for a long time than running a density functional theory (DFT) calculation on a classical computer, which is even more costly than an AI prediction. Having said so, a quantum computer would be a fantastic source of *numerically exact* (within a basis set error) training data for both DFT/wave function methods and AI approaches. Therefore, I think of these tools as a Russian doll of complexity rather than any one eclipsing the others. Check out our recent work on where the quantum computing field in chemistry is and why we think it will be useful in the long run.[1]

1 P. Schleich, *et al.*, *arXiv*, 2024, preprint, arXiv:2401.09268, DOI: **10.48550/arXiv.2401.09268**.

Varinia Bernales communicated: Alán, thank you for an excellent paper (**https://doi.org/10.1039/d4fd00153b**) and a beautiful talk. During your presentation at the *Faraday Discussion*, you mentioned that our next steps should focus on developing AI scientists capable of generating hypotheses and running experiments alongside human researchers. My question is, why should the scientific community focus on creating these AI agents when we might instead concentrate on improving educational systems and providing more development opportunities across the globe? By doing so, we could enable more people to ideate, create, and produce new ways of thinking and new technologies. Is this a sign that we are losing hope in the potential of human innovation? Additionally, I wonder about the ethical concerns involved in developing such an AI scientist and what safety guardrails would be necessary.

Alán Aspuru-Guzik communicated in reply: Thank you for your question and your comments on the talk.

Let me address your question in parts. Firstly, increasing access to educational and social opportunities can level the playing field and enable many more humans to become scientists. As a Latin American, I find this pretty acute, given the lack of resources in our region. Both you and I have commented about the equity of access to science in the recent *Chemical Engineering News* Trailblazers issue about Latin American Scientists.[1–3] I strongly believe that we should work together to enable as many young minds as possible to become future scientists.

On the other hand, scientific progress needs to be made faster. Eroom's law shows that developing a drug can take more than a billion dollars and about a decade of work.[4] This prevents millions and millions of people from having access to potentially life-changing drugs. The Arctic is melting at a frantic pace,[5] and we are losing biodiversity in the Amazon due to deforestation. Chemistry plays a central role in developing solutions for these challenges. If AI-driven labs

can help accelerate progress in these fields, humanity must work on this subject to ensure we find novel solutions to these global problems in time.

Finally, I fully agree that these AI-driven labs' ethics and safety are paramount. We want to ensure we don't create unintended consequences or harm human beings. I just returned from delivering a talk at a university, where I heard that AI-designed conditions and a careless error led to a lab accident. We need to be mindful that safety has to be built into our systems from the get-go and that we take a step to pause and think about our research consequences while simultaneously accelerating the pace of research. A daunting paradox!

1 https://cen.acs.org/careers/diversity/The-Revolucion-Revolucao-will-not-be-televised/102/i29.
2 https://cen.acs.org/safety/Varinia-Bernales-develops-safer-chemicals-while-making-science-more-equitable/102/i29.
3 https://cen.acs.org/physical-chemistry/computational-chemistry/For-Alan-Aspuru-Guzik-boundaries-between-chemistry-AI-and-robotics-do-not-exist/102/i29.
4 J. Scannell, *et al.*, *Nat. Rev. Drug Discovery*, 2012, **11**, 191–200, DOI: 10.1038/nrd3681.
5 N. Wunderling, *et al.*, *Nat. Commun.*, 2020, **11**, 5177, DOI: 10.1038/s41467-020-18934-3.

Roohollah Hafizi opened discussion of the paper by Chris J. Pickard: In Section III of your paper (https://doi.org/10.1039/d4fd00134f), the random structure generator is referred to as a 'generative model' in machine-learning terminology. From a technical standpoint, a generative model typically identifies patterns in the data and uses that information to generate new content similar to the original data. Do you think that the "random structure generator" aligns with this definition of a generative model?

Chris J. Pickard answered: Absolutely. We build distributions of what we call "random sensible structures". Whether the structures are sensible or not depend on how well they conform to the patterns in experimental data and theoretical facts that have been learned over the last century or so. We stochastically choose a reasonable unit cell volume for the structure – it is observed that structures do not vary that greatly in their densities at a given pressure. We make sure atoms do not sit on top of each other – we have never seen that in experimental crystal structures. It is observed that different elements have different sizes, or atomic radii, and this determines the range of randomly chosen interatomic spacings. We know that some elements do not preferentially sit next to either others, or themselves. This can be encoded into an inter-species minimum distance matrix – MINSEP in the *ab initio* random structure searching (AIRSS) buildcell syntax. It is known that structure might be hierarchically built, and so we can generate molecular crystals by randomly placing molecules into the unit cell so that they don't overlap, as opposed to their individual atoms. This extends to fragments in inorganic materials and combining the fragments with molecules for hybrid structures. As described in Section VI of the paper (https://doi.org/10.1039/d4fd00134f), these parameters can be directly measured from a given structure and used to generate new structures that are geometrically very similar to the seed structure, and likely low lying on the energy landscape. This is taken a step further in Section VII, where we measure environment vectors, and use those to generate new structures with similar environment vectors. I later discuss the close relationship of this method to modern generative diffusion-based methods.

Roohollah Hafizi asked: You mentioned that your approach tends to guide the simulation towards regions of configurational space presumed to be "good". In your opinion, should structure prediction methods prioritize biasing the search towards known favorable regions, or should they place more emphasis on exploring less visited areas?

Chris J. Pickard responded: This entirely depends on the nature of the search you a performing and is an example of the "exploration/exploitation" trade-off. If you are carrying out a polymorph sweep for a composition for which you know the crystal structure of one or two examples, it makes sense to perform a very focused search using parameters measured from those known structures. The random structures will be geometrically like those seed structures, which you know are relatively low in energy, and you will explore the nearby space of possible configurations. However, if you are trying to understand the chemistry of, for example, the core of a giant planet, you cannot assume that our accumulated chemical knowledge is useful, and you should relax your constraints. However, once you have performed a few, less biased, AIRSS searches you will start to observe the expected densities, and interatomic separations, and subsequent searches can exploit those. It should be noted that a key feature of AIRSS is that as far as possible structures are generated to be uniformly distributed within the distributions defined by given, and clearly stated, constraints. We want to make sure that we probe the edges of the distributions, as opposed to "regressing to the mean" as can occur in approaches that target average values. This increases the chance of genuine discovery.

Heather J. Kulik enquired: As the unit cells get more complex, is there way to constrain sampling to preserve molecular units, such as in cases where there are large but stable organic components such as benzene rings or methyl groups?

Chris J. Pickard replied: Yes, the AIRSS buildcell syntax is designed to permit the packing of connected units, which might be user supplied molecules, or structural fragments. The AIRSS cryan code provides algorithms to extract connected units from a provided crystal structure. There are two approaches, the first based on graphs formed by connecting atoms within a given distance of each other (which may be species dependent), and an analysis of the spectral properties of the Laplacian matrix. The second approach performs community detection, as described in ref. 1.

1 S. E. Ahnert, W. P. Grant and C. J. Pickard, Revealing and exploiting hierarchical material structure through complex atomic networks, *npj Comput. Mater.*, 2017, **3**, 35.

Heather J. Kulik asked: How do you decide which things are connected in your algorithm, is it based on distances?

Chris J. Pickard responded: Yes, distances are used to determine connectivity. Internal species dependent defaults are supplied, or the threshold distances can be chosen by the user.

Bianca Pasca requested: In your work, you have mentioned using symmetrised molecular dynamics and carrying out these simulations on your in-house code, ramble. Can you elaborate on how this is achieved?

Chris J. Pickard replied: The ramble molecular dynamics code implements symmetry constraints by symmetrising the velocities (which determine the update steps) in the same way that forces and stresses are symmetrised for geometry optimisation. This ensures that the symmetry operations of the initial space group are adhered to throughout the simulation.

Aron Walsh asked: Beyond crystal structure prediction, one of the exciting aspects of generative artificial intelligence models is the ability to condition against properties. Can your methods be extended towards navigating property landscapes?

Chris J. Pickard answered: In my research I prefer to focus on computational "discovery" as opposed to direct "design" of new materials. I would first identify novel phases that are thermodynamically, and more recently kinetically, stable, or moderately metastable, and therefore feasible targets for synthesis. I would then screen those structures for candidates with desirable properties. This is the approach that we took that led to the computational discovery of Mg_2IrH_6 as a candidate high temperature superconductor with a predicted T_c of 160 K at ambient pressures.[1]

It is possible to modify the energy landscape, or search algorithms, to favour materials with designed properties. An example is the search for dense materials (which might be also superhard). By adding a term proportional to the volume of the system to the internal energy, low volume (dense) structures are favoured. The constant of proportionality is identical to pressure in the enthalpy, $H = U + pV$. A limitation of this method is evident – experimental high-pressure synthesis must contend with the recoverability of the material to ambient conditions. In most cases high pressure phases are only stable under an applied external pressure. Care must be taken to ensure that the design inspired penalty function does not significantly move the lowest lying local minima of the energy landscape. It is not clear how to achieve this in general. Many of the designed structures are expected to be high lying metastable states, if dynamically stable at all. For evolutionary algorithms the fitness function can be adjusted to permit the imposition of designed properties.

1 K. Dolui, L. J. Conway, C. Heil, T. A. Strobel, R. P. Prasankumar and C. J. Pickard, *Phys. Rev. Lett.*, 2024, **132**(16), 166001.

Graeme M. Day requested: You commented on advice that you would give to early career researchers: to look for problems that are on the edge of being possible. Can you comment on the current problems at the limits of what is possible with crystal structure prediction that people entering this field should look at tackling?

Chris J. Pickard responded: The general principles of crystal structure prediction are well established, but of course all methods will struggle with

sufficiently complex systems, where an "exponential wall" of complexity is encountered. There is always scope for pushing up against that wall. We might dream of predicting protein crystal structures, with the conformational folding included. But that is a way away. I would suggest looking beyond perfect crystal structure prediction – tackling the challenges of predicting defects, disorder and microstructure. While there has been some limited progress towards this from first principles (see ref. 1 and 2) the arrival of machine-learned interatomic potentials (MLIPs) opens new opportunities, and it is likely that this direction develops rapidly in the coming years.

1 A. J. Morris, C. P. Grey, R. J. Needs and C. J. Pickard, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **84**(22), 224106.
2 G. Schusteritsch and C. J. Pickard, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **90**(3), 035424.

Venkat Kapil asked: Have you considered doing a formal analysis of the time complexity of the standard and 'hot' random structure search methods?

Chris J. Pickard replied: I have considered this, but with no conclusion so far. But I agree that this would be a very interesting direction to explore.

Christian Kuttner enquired: What are the specific advantages of the ephemeral data-derived potential (EDDP) distance-based approach over previous methods in generating candidate structures, and how does it impact the discovery of novel materials?

Chris J. Pickard answered: This is a good question. The EDDP distance-based approach will generate structures that are likely to be even more like the target, known, structures than the traditional AIRSS buildcell algorithm, particularly for open covalently bonded systems. This is expected to be beneficial for a polymorph sweep, but possibly detrimental to the chances of genuine discovery. The EDDP distance-based approach is an extreme example of favouring exploitation of known information, at the likely cost of reduced exploration.

Wenhao Sun asked: Hot AIRSS certainly appears to be helping to find the crystal structures for boron, despite what I wrote in my last question about the complexity of the possibility of charge disproportionation in boron. I'm wondering if you think hot AIRSS is guiding the system towards the proper physics relevant for crystal structure selection, or do you think that the significantly reduced cost of simulation is just allowing you to explore a broader structure space? (Or both?)

Chris J. Pickard responded: I believe that the physics is being correctly described, for a sufficiently well trained EDDP MLIP, because the temperatures that provide the greatest acceleration of the hot-AIRSS search are at slightly below the experimentally known melting temperature of boron. This suggests that the energy barriers to bond breaking, making and rearrangement are realistically described. In ref. 1, we showed that EDDPs give a good description of a variety of properties, including melting temperatures, phase diagrams and diffusion rates.

1 P. T. Salzbrenner, S. H. Joo, L. J. Conway, P. I. C. Cooke, B. Zhu, M. P. Matraszek, W. C. Witt and C. J. Pickard, Developments and further applications of ephemeral data derived potentials, *J. Chem. Phys.*, 2023, **159**(14), 144801.

Matthew R. Ryder communicated: With the growing integration of data-driven methods and machine learning in computational chemistry, how do you see the field evolving to include more accurate treatments of disorder and thermodynamics in materials discovery? What are the main barriers to this?

Chris J. Pickard communicated in reply: As exploited in hot-AIRSS, the dramatic acceleration of both molecular dynamics and structure prediction by machine learned potentials with close to DFT accuracy opens up the possibility of routinely introducing the effects of entropy into our simulations. We are seeing a rapid recognition of this in many research directions. In some ways the barriers are low – compared to first principles methods. However, a concern is that these machine learned potentials can't exceed, and will rarely meet, the accuracy and robustness of density functional theory. Where in an electronic structure calculation you can simply recommend that a new user converge their results with respect to the key computational parameters, ensuring reliable and reproducible results from approximate machine learned potentials will require care.

Wenhao Sun opened discussion of the paper by Christopher M. Collins: First, I want to say that I was very impressed with the scope of the systems you've selected. I don't think I've ever seen such complicated materials being tackled by crystal structure prediction methods before, especially this 7 component Si–Al–B–Fe–Na–O–F phase. I just want to clarify my understanding, is FUSE able to actually get the correct crystal structure, or are you just checking to see how low in energy various methods get?

Christopher M. Collins replied: Thank you! All of the probe structures which we obtained during the study have been provided as part of the GitHub repository where FUSE is hosted. Among the known phases, we did only obtain the $Ca_3Ti_2O_7$ phase. We specifically chose systems which we had been confident that we could not obtain, so as to get as much useful information as possible as to how each of the different setups were performing. There were some examples from the original FUSE paper[1] which we started studying, but proved to be too trivial – if the search can locate the global minimum quickly, we do not learn very much.

Since the paper (**https://doi.org/10.1039/d4fd00094c**) was completed, we have additionally completed runs which obtain the correct structures for the $CoAs_2$ and Cu_7S_4 phases. For the three Li containing compositions, we have been unable to improve on the minima included in the paper, and are increasingly confident that these are the global minima for these compositions. Based on the experiments in this paper, we have learned a lot about how this new implementation of FUSE is working and have made several modifications to the code, and will be conducting another round of calculation benchmarking (possibly with an expanded set of compositions) in the near future and will provide results alongside a future release of the code.

1 C. Collins, G. R. Darling and M. J. Rosseinsky, *Faraday Discuss.*, 2018, **211**, 117–131, DOI: **10.1039/c8fd00045j**

Graeme M. Day asked: You commented that you see diversity of structures from the crystal structure prediction using generative models. Diversity of structures is important for finding novel materials in searches. Have you tried to quantify diversity of structures in generative *vs.* random searching? And to follow on from this, if the basin hopping is seen to lead to a diverse final set, do you think that the optimisation method, *e.g.* basin hopping *vs.* simulated annealing *vs.* genetic algorithm, will have a more important role when starting from generative model structures than random structures? My experience when starting from random structures is that the optimisation method has a small influence.

Christopher M. Collins responded: Not successfully, previously any attempt we have made to make or use a metric to measure this, we have got as far as being able to tell that two structures are different, but without placing a useful numerical value on the difference. The only real exception to this is to look at the range of energies among the generated structures, where generally these models do give a wide range. This is something that we are keeping under review, as we would like to have metrics that we can use to describe the diversity of generated crystal structures, not just their range of energies.

Regarding the second part of the question, for the types of materials we study, we have found that basin hopping makes a huge difference in how effectively we are able to locate low energy structures, certainly *vs.* only optimising random structures. This is borne out in the results of the paper (**https://doi.org/10.1039/d4fd00094c**), where in 58 out of the 66 experiments that we performed using generative ML models, the basin hopping improved on the energy obtained from the ML model, generally in our application, the same is true of using basin hopping *vs.* random structures. Table 6 in our paper additionally illustrates that the energy gain from using basin hopping for most of the benchmarks is substantial, this point is further illustrated in Fig. 7 in the paper, where we show what the impact is of this in terms of the obtained crystal structure. I do expect, that as generative ML models for structure generation improve, the difference between the initial energy and the energy obtained by the subsequent basin hopping will narrow. We have previously investigated the use of genetic algorithms in the context of FUSE, and have found it to be far less effective than basin hopping. This is surprising given that within the FUSE code, it is easy to describe a crystal structure as a 1D sequence of the blocks from which structures are assembled, which we would imagine would be ideally suited to being optimised *via* a genetic algorithm.

I would add to this that the use of random structures in crystal structure prediction is still important: creating a population which is well sampled initially seems to reduce the likelihood of searches starting (and getting stuck in) high energy areas of the potential energy surface and provide an effective way of escaping being trapped in local minima.

Roohollah Hafizi enquired: Regarding crystal structure prediction through random sampling, have you benchmarked your AI-generated structures to assess whether they demonstrate greater efficiency compared to evolutionary algorithm implementations, such as USPEX? What was your sampling size?

Christopher M. Collins replied: No. We have always felt that running benchmarks ourselves against other structure prediction codes is not beneficial. Any tests which we perform, will leave our own methods at an unfair advantage: as we have created the code we have a good understanding as to how we can run it most effectively, we do not have that level of knowledge for running codes produced by other research groups, and could readily be offered as a reasonable explanation for results which appear to show one method as superior to another. We would however, encourage other code developers to test codes on benchmark systems such as those we have included in the paper (**https://doi.org/10.1039/d4fd00094c**). As our own codes develop, we will expand this set to include additional challenging known materials and increase the number of compositions for which single phases are not known to exist. Including systems without known compounds, reduces the likelihood that high-performance when using ML models is purely because of the mode being able to recall known crystal structures. We will consider an appropriate way to disseminate this expanded set and our results as FUSE continues to develop.

Christian Kuttner communicated: Looking at avenues to enhance the efficiency of crystal structure prediction, how exactly do the generative ML models improve the starting population for heuristic algorithms compared to traditional methods?

Christopher M. Collins communicated in reply: From the work in our paper (**https://doi.org/10.1039/d4fd00094c**), we compare against FUSE's original structure generation method, which is a based on randomly allocating atoms to a predefined series of positions within the submodule motifs that FUSE uses. We observe two main differences between our experiments which rely on the original generator and our ML model:

(i) When using the ML generation, the compute time to optimise a structure to a local minimum is faster (this is irrespective of whether or not we use a machine learnt interatomic potential during this process or not).

(ii) In most experiments, when using ML structure generation, the resulting structures are lower in energy. As the experiments were conducted within a fixed amount of compute time, I suspect that the second observation is a result of the first. So this would suggest that the main improvement over our classical method in the initial population is from producing structures which are closer to energetic minima and so are more cost effective to compute, further elaboration would require further experiments to disentangle.

This, of course, only considers the comparison between FUSE's original generator with one ML model. Since the paper was completed, we have started to investigate the use of other structure generators, in the form of other ML models, and the AIRSS method. Our current hypothesis is that we are unlikely to arrive at one structure generation method which is superior to all others, and so one option for the future will be to use multiple structure generation models concurrently, for a given composition of interest, and combine the results to form the initial pool of structures for the structure prediction search. This is something that the current implementation of FUSE is setup to do; in future updates of the code, we will implement additional structure generators ourselves and write

a guide for how other users could implement their own generation models/methods.

Matthew R. Ryder communicated: In your hybrid approach combining heuristic and generative machine learning, how do you address the challenges of incorporating real-world complexities, such as noncovalent interactions and dynamic disorder, into structure prediction?

Christopher M. Collins communicated in response: I think that the general answer to your question is that we do not. The reason for this, is that many of these complexities are very difficult to model, and generally, do not have a one size fits all approach that can be applied at scale (currently!). Instead, we use the ordered models that we generate to probe the chemical space, to find materials which are likely to yield new compounds. Because of the constraints that are necessary to make such calculations tractable, details such as disorder, whatever the cause, are unlikely to be captured by the models. This answer is also linked to another comment, that another reason for using this "probe structure" approach[1] is that we are unlikely to ever be able to guarantee that we truly have predicted the ground state structure, or exact composition. This use of structure prediction to guide towards new compounds does prove very effective in practice, then once we have new compounds synthesised within the lab, we study them experimentally, with further computational support where required.

Other examples of where we have used probe structures as guidance include ref. 2 and 3.

In all these cases, the experimentally isolated materials contain substantial levels of disorder and structural complexity, but our "probe structure" approach to exploring chemical space is still able to locate them. While all these examples are using structure prediction methods which pre-date the one which I presented at the conference, the same points are valid, while the use of machine learning models is increasing the speed with which we can make predictions, they are not currently overcoming complexities referred to in your question.

1 C. Collins, M. S. Dyer, M. J. Pitcher, G. F. S. Whitehead, M. Zanella, P. Mandal, J. B. Claridge, G. R. Darling and M. J. Rosseinsky, *Nature*, 2017, **546**, 280–284, DOI: **10.1038/nature22374**.
2 C. M. Collins, L. M. Daniels, Q. Gibson, M. W. Gaultois, M. Moran, R. Feetham, M. J. Pitcher, M. S. Dyer, C. Delacotte, M. Zanella, C. A. Murray, G. Glodan, O. Pérez, D. Pelloquin, T. D. Manning, J. Alaria, G. R. Darling, J. B. Claridge and M. J. Rosseinsky, *Angew. Chem., Int. Ed.*, 2021, **60**, 16457, DOI: **10.1002/anie.202102073**.
3 G. Han, A. Vasylenko, L. M. Daniels, C. M. Collins, L. Corti, R. Chen, H. Niu, T. D. Manning, D. Antypov, M. S. Dyer, J. Lim, M. Zanella, M. Sonni, M. Bahri, H. Jo, Y. Dang, C. M. Robertson, F. Blanc, L. J. Hardwick, N. D. Browning, J. B. Claridge and M. J. Rosseinsky, *Science*, 2024, **383**, 739–745, DOI: **10.1126/science.adh5115**.

Filip T. Szczypiński opened discussion of the paper by Brett M. Savoie: Most of the properties in your large property model were calculated using GFN-xTB. Those would typically relate back to the electron density and be calculated with established operators, meaning that there is substantial correlation (or even colinearity) of the properties in your model. How do you treat such highly correlated features and could that be the origin of the property redundancy resulting from your model that effectively reduces the dimensionality of the property space?

Brett M. Savoie answered: Great point. The redundancy of the properties is what we are counting on so that the property to graph distribution mapping becomes tractable to learn, but it is difficult to judge how many properties will be required *a priori*. The fact that most of these properties were calculated with GFN2-xTB creates the possibility that there are some spurious property correlations that the model might be learning that are unphysical. The property masking task should help with this, but ultimately better and more general property datasets are the answer.

Filip T. Szczypiński questioned: You classify molecules as distinct if they exhibit an effective separation of 1% in at least one property. The properties are calculated at GFN-xTB level for a single conformer obtained with Auto3D. Many properties will carry an error greater than that cutoff, especially for properties that strongly depend on the conformation, such as a dipole moment. Could the masking that you introduced in the discussion help to alleviate the problems related to the quality and uncertainty within the underlying data?

Brett M. Savoie replied: We agree, the errors associated with the underlying properties affect distinguishability. The 1% difference was chosen for convenience, but property specific uncertainties would be better in general. The property uncertainty also impacts how the user should interpret the molecules generated by the model. If the user is sampling from an uncertain property distribution then they would probably benefit from oversampling structures and then ranking using an independently trained property predictor. For example, if your target properties were high synthesizability and low toxicity, then you might oversample molecules using the large property model (LPM) with your targeted property profile then filter the structures using a more accurate predictor (*e.g.*, retrosynthesis planner or a specific toxicity regressor).

Niamh Hickey enquired: Prior to running your model, were there any particular properties (or a combination of properties) that you thought might be instrumental in defining the structures but ended up not being so?

Brett M. Savoie answered: Thanks for the question. Honestly, we didn't really dwell on it, because I assumed that we would be deeply in the data scarce regime and so every property would count. I still think we are somewhat in that regime with respect to inverse design across broader chemical space.

Shubham Vishnoi asked: How do you handle conformation-dependent properties in your generative models for molecules, given that the same molecule can adopt different conformations leading to varying properties? Are there specific strategies employed in the large property models you describe to account for these variations, or are there approaches that could be implemented to enhance their predictive accuracy?

Brett M. Savoie replied: Terrific point. Formally speaking, every graph2property problem is underdefined because all properties are conformationally dependent. Nevertheless, models don't seem to get confused, so what is going on here? The answer is that deep models trained on graph2property tasks have an

inductive bias to predict properties associated with the conformer generator routine of the training data. In most cases, this is the lowest energy conformer discovered using a particular algorithm. The inductive bias makes the problem one to one and well defined, but it is supplied indirectly in the way that the training data is curated. That's a much longer answer than is required here, but I don't think this point is appreciated in proportion to its importance. The short answer to your question is that we handle conformation dependent properties by only taking the values associated with the lowest energy conformer. Other choices would be to Boltzmann average the properties (probably more accurate) or to sample the property vectors during training using an uncertainty estimated from the conformational distribution. This might have the indirect benefit of supplying the LPM during training with the uncertainties associated with the property vectors.

Shubham Vishnoi queried: Can the large property models pipeline used for the inverse design of molecules be applied to crystal structure prediction? If not, what adjustments would be necessary to make it feasible, and do you think it might become possible in the future?

Brett M. Savoie responded: Terrific question. There isn't a fundamental obstacle to applying this approach to crystals. However, you would need to use a different decoder architecture with a more sophisticated syntax capable of describing crystal structures (SMILES, used here, only expresses molecular graphs with limited geometric annotation).

Annabel Eardley-Brunt asked: On the topic of data masking, this is a really interesting study and has lots of implications for studying data scarce space, and for areas of science that have problems with data scarcity which make using AI methods a challenge. How do you think that the masking results will hold up as the training data size changes, particularly for smaller datasets? As many of the property variables used are intrinsically related, do you think that the model is able to compensate for missing variables as it has truly learned the chemistry, and so is able to make calculations and approximations based on the variables that are present? Or do you think that the model is simply finding relationships between the variables and compensating for it that way. Or are these two points actually the same thing, and so the concept of a model being able to 'think' like a chemist is not as far away as we imagine?

Brett M. Savoie answered: Great question! My perspective is that learning the relationships between properties, which is what the masking task is meant to promote, is functionally equivalent to the model learning chemistry and what we think of as structure–function relationships. Interpreting what the model is actually learning and how it is making particular inferences are great research questions. The question of how these models will perform in the data scarce regime is probably the most important for future work. Here we've shown that this formulation is learnable and have some proof of principle case studies, but we have a lot of things planned. One way to test this is to retrain the models while imposing data scarcity on subsets of properties and then monitoring the accuracy *vs.* size learning curves for those properties.

Alán Aspuru-Guzik said: This is a question and a comment. In the field of quantum mechanics, there is a useful concept (originally spearheaded by the ideas of Ernest Schrödinger on wavefunctions and measurement) called quantum tomography. In quantum tomography, one can show that for a finite quantum system, there is a finite set of measurements that fully describe it. My group and I have applied it to fields such as molecular spectroscopy.[1] I think there is a super cool analogy between your work and quantum state tomography. As you work towards "measuring" using properties, the "wave function" (discrete description of the molecule) begins to emerge. I think further connections between these two apparently disconnected topics could emerge such as the very popular recent ideas of shadow tomography.[2]

1 J. Yuen-Zhou, J. J. Krich, M. Mohseni and A. Aspuru-Guzik, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 17615–17620, DOI: **10.1073/pnas.1110642108**.
2 S. Aaronson, *arXiv*, 2018, preprint, arXiv:1711.01053, DOI: **10.48550/arXiv.1711.01053**.

Brett M. Savoie responded: Thanks Alán, I love this comment. This problem of "what constitutes sufficient information to determine [fill in the blank]" comes up again and again. Outside of controlled problems where you can fully enumerate the degrees of freedom, my understanding is that it is very difficult to pin down what constitutes sufficient information for an arbitrary prediction task. Practical limits can be established by training and testing regressors as a function of supplied information, but this is expensive and not terribly satisfying.

Alán Aspuru-Guzik asked: Wonderful talk. I love the idea of going from properties to molecules in this architecture. Your paper (**https://doi.org/10.1039/d4fd00113c**) is a very novel contribution to the field. I wonder if you considered using a non-linear set of combinations for your model? I am inspired by autoencoders, for example, where the most significant dimensions of the latent space represent a very compact combination of descriptors.[1]

1 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276, DOI: **10.1021/acscentsci.7b00572**.

Brett M. Savoie responded: Alán, thanks for the generous feedback. I'm optimistic that in a future iteration we will be able to find an irreducible set of non-linear property combinations that fully determine the structures. Indeed, the decoder in our current architecture is already working off of a tensorial fingerprint generated by the encoder from a non-linear combination of the properties. To address your point we would need to carefully tune the dimensions of this latent space but we haven't done that yet.

Magdalena Lederbauer asked: The question of getting from a property to a graph/chemical compound is particularly interesting in a chemistry setting. I am curious about taking this question one step further: in a real-world scenario, chemists measure data in the form of experimental spectra (MS, IR, NMR, …). Can we use this approach for structure elucidation as well – *i.e.*, derive a (latent) property vector from spectra? Possibly both the noise contained in an experimental spectrum and data availability will be limitations here.

Brett M. Savoie answered: Terrific question. Indeed, there is nothing particularly fundamental about using scalar or categorical property vectors. Spectra can be considered vectorial or tensorial properties of a molecule. Looking towards the future, I would guess that we will see more multimodal reasoning across scalar and non-scalar molecular "properties".

Steven Torrisi said: I greatly enjoyed reading this paper. The idea of trying to find uniquely identifying properties of a molecule – as a previous questioner noted – would be a great complement to the work of *e.g.* spectroscopists and other characterization experts, as it might assist in optimal design of experiments (so as to maximize knowledge about properties that would be maximally identifying/discriminating towards a molecule or system of interest).

The passage in your manuscript (**https://doi.org/10.1039/d4fd00113c**) where you compared the distribution of molecules and properties to the distribution of ideal gas particles in a volume got me thinking about how to get a sense for how 'low-dimensional' the manifold you are working on is.

Perhaps another way to make sense of your findings is to compare the ~20 properties you are considering against hypothetical lower-dimensional descriptors of each molecule. Suppose that we could describe molecules where the list of properties come from a series of 'yes/no' questions, where the answers are stored as 0/1 (or True/False; basically, boolean variables). For 1.3 million molecules, $\log_2(1\,300\,000) = 20.31$. Since the total number of unique 0/1 strings for a set of N bits scales as 2^N, in theory 21 well-chosen True/False statements about the molecules in interest could uniquely identify each and every molecule in the dataset (examples might include the presence/absence of a particular motif; a particular element, or so on). To summarize, that would be the size of the space needed, for the lowest dimensional descriptors possible, to uniquely identify all molecules in the set.

My next thought goes to how we might imagine a strategy to map from *e.g.* spectral data to try to identify molecules in this library, as well as to get a sense for the intrinsic challenge of working with these low-dimensional descriptors. To take the hypothetical further, suppose we had an independent model for each of those questions mapping from *e.g.* some observable to 0 or 1. Suppose, say, mapping from a spectrum to 0/1 for one of the properties. If we assumed the classification accuracy was the same for all models, the odds at least one of these models makes a mistake would be described by the geometric distribution – accuracy in the >96% range would be required to ensure all 21 models correctly give the right answer more than half the time. Of course, this situation in the manuscript is quite different, including that: you are not making independent models for each property, the property space you're using is higher-dimensional than a set of booleans, and the goal is not perfect identification (top-1 identification in your paper, where you obtain an impressive ~35% accuracy given the challenge of the task. Additionally, a few wrong binary classifications could be tolerated along a 'top-10' style reporting).

I am very curious about what may come from exploring the empirical observation that a subset of ~10 parameters works well for practical purposes – such as colinearity testing of the properties you're using. I'd welcome any thoughts you have about the above, such as how you might exploit combinations of values of

various properties to help identify discriminating sets of properties for identifying particular molecules for future users.

Brett M. Savoie responded: Great comments. The bit analysis showing that 21 well-chosen questions should be sufficient to encode the 1.3m library size studied here is consistent with our thinking. With the definitions in the manuscript (**https://doi.org/10.1039/d4fd00113c**) the relevant base would be 100 because we assume that a 1% difference in properties is numerically distinguishable. Other choices are possible. The second point about the accuracy necessary to reproduce a growing list of properties is really important. We mostly reported errors with respect to getting all of the properties right, but in practice the user would only care about reproducing a subset of properties. The achievable accuracy for reproducing all properties should monotonically decrease as the number of properties being specified increases.

Joshua Cheung enquired: In your results, was there any difference in model performance when predicting the structure of an organic compound compared to an inorganic compound? For example, were more or fewer properties required to distinguish an organic *vs.* inorganic compound?

Brett M. Savoie answered: We're very interested in this question. For the current case studies everything was restricted to CHONFCl training structures, so we are unable to compare at this point.

Joanna Grundy questioned: Given the large space of possibilities given 22 binary variables, over 4 million, and that even with only 10 variables it didn't matter which ones, is it possible the system has just rote learned the space rather than learned a chemical manifold? See ref. 1 for examples. A way to check this would be to use a random vector instead of the features just to see if it can perform as well.

1 C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, *arXiv*, 2017, preprint, arXiv:1611.03530, DOI: **10.48550/arXiv.1611.03530**.

Brett M. Savoie replied: The suggestion to test on random vectors – and more generally sampling from other property distributions – is an excellent one. We did not try that here. All testing property vectors were drawn from a distribution of unseen testing molecules that nonetheless corresponded to the real molecules. Put differently, the question is what happens when we give the LPMs unphysical property combinations? We'll see in future work. Regarding whether rote memorization is occurring, the current case studies are sufficient to falsify this. All results are for unseen testing sets. Additionally, the reconstruction tasks are extremely challenging. The nearest trivial benchmark would be a model that generated a random constitution isomer consistent with the elemental formula. For the typical inference task, this would be a space of tens of thousands of structures and such a model's performance would approach 0%.

Roohollah Hafizi asked: When the molecular graph is mapped to a list of properties, the length of output layer remains fixed. However, when this process is

reversed, the output length can vary, as molecules of different sizes may exhibit similar properties. Could you clarify whether your model is designed to generate molecules of varying sizes?

Brett M. Savoie answered: Great question. The decoder predicts token by token and the graph terminates once the end token is predicted. This makes the molecular size technically open ended. However, there is still a context window for the decoder that imposes a limit on how much of the molecular graph is in view for the decoding. I don't have the number in front of me, but this is typically very large $O(100–1000)$ tokens, but it would probably be a problem for macromolecules.

Gillian James said: I have a question regarding the masking scheme: certain properties may have empirical or analytical relationships between one another which may inform structure prediction in special ways. For example, materials with a high bulk modulus will be more likely to have a small entropy, and may thereby have higher packing, be stabilized at low temperature, *etc.* For these properties that have well-established or loose relationships between one another, I would imagine that the more interconnected the property, the higher the impact of masking this property would be on the structure-prediction accuracy. In other words, are some properties more informative than others for structure prediction and have you thought about ranking the relative 'informativeness' of different properties?

Brett M. Savoie responded: Thanks for the example. This is in line with our thinking that there are (as yet undetermined) basis sets of optimally informative properties. The current case studies don't directly speak to this and so we are reserving judgement on the specific informativeness of different properties (or their combinations). We'll just add that the evaluation case studies tend to be quite expensive owing to the high novelty of new structures during generation. This means that every case study involves the evaluation of properties for many new molecules.

Mark Goulding asked: In your paper (**https://doi.org/10.1039/d4fd00113c**) you scraped 1.3m materials datasets from PubChem from across the spectrum of properties. You state that data integrity of these datasets is beyond the scope of a *Faraday Discussion* and that they are regarded as "ground truth" for your study. How concerned are you about this for the accuracy and direction of your models?

How much better would you expect your models to be if they were trained with an experimental structure/property dataset where test conditions are known and identical for every material + data integrity is high, giving an absolute relationship?

Brett M. Savoie responded: Excellent question. The LPM architecture will only learn the conditional distribution $P(G,\text{properties})$ associated with the dataset it is trained on. This means that it will reflect the biases and inaccuracies of the underlying training data. One symptom of inaccurate data would be broader structural distributions as the model returns structures that don't match the

specified properties. A potential advantage of better data is that you might need less of it. Nevertheless, we are still in the strong scaling regime in these experiments and so we expect the more data the better at this stage even if it is of middling accuracy.

Graeme M. Day communicated: I have a question about masking in the large property model training. For scalar inputs, this was performed by assigning the mean value from the training distribution, and a special token for class-based inputs. I wondered if the way that masking is performed for scalar inputs has an influence. By assigning the mean value, the model sees what looks like a realistic value from the distribution during training. Would it make a difference to training and performance if an unrealistic value was used for masking, to let the model learn more easily to ignore that property? Or does the value used in masking have little influence?

Brett M. Savoie communicated in reply: You are right, this isn't the only choice. In general, you wouldn't want to mask with a dummy scalar that was a silly number, because the linear embedding network will probably get confused by trying to sensibly embed the distribution of unmasked values and this singular out of distribution scalar with the same set of weights. Our implementation of your suggestion would look like having a separate embedding layer for the mask value for each scalar, which would treat them the same as the masked categorical variables. The reason we went with the mean was that we thought that it might be somewhat informative, given that the user would have access to the training distribution for that property. So why not potentially use that information rather than using a generic information free masking value? The contrary argument is that the mean may be unphysical to combine with the rest of the unmasked values and so this might actually confuse the model. We got stable training for the choice of the mean, but the other implementation has yet to be tested.

Ksenia R. Briling communicated: As I understand, the 80 constraints/trivial properties are present in both training and inference (section 3.2; **https://doi.org/10.1039/d4fd00113c**). Have you compared the performance to a case without using these constraints whatsoever? It seems that *e.g.* the high formula reconstruction accuracy comes from the constraints that (partially) encode the formula in the query. In practice, is imposing constraints beneficial compared to filtering out the molecules that do not satisfy "design constraints" from an unconstrained prediction?

Brett M. Savoie communicated in response: This is a great point. The constraints help stabilize the training and almost certainly contribute to the high molecular formula reconstruction accuracy. With all constraints fully specified there are still thousands of consistent constitutional isomers for the size of molecules we are studying. So for the unmasked situation, you could interpret the top-1 reconstruction as an accuracy with respect to the possible constitutional isomers. For the masking cases it is different. We have yet to do the test where we preferentially mask most or all of the constraints at a higher rate than the properties. Our interpretation is that the constraints help stabilize the early training by localizing the model on the lower dimensional physical distribution of

molecules, but that at later fine-tuning stages these constraints could be relaxed. Regarding "In practice, is imposing constraints beneficial compared to filtering out the molecules that do not satisfy "design constraints" from an unconstrained prediction?" This is a good question! This will entirely depend on the inference cost and the cost of filtering. For some property combinations it might take many samples to get structures that don't have, say, a carboxylic acid. But with the constraint specified you can preferentially sample structures without it.

Ksenia R. Briling communicated: How is property reproduction for new molecules evaluated? If all the properties are mapped to 0–100% interval, could it be possible that a $\pm x$% hit performance measure is affected by large/small variation across the dataset?

Brett M. Savoie communicated in reply: Evaluation is straightforward but expensive. We generate molecules with the LPM and a given property vector, then we take those molecules and run the property calculations on them to generate a ground truth property vector for each molecule. The errors in these properties are reported on a percent basis with respect to the inference property vector. This is expensive because the LPM is generally generating completely new molecules that we have to calculate properties for. The choice to evaluate everything on a percent normalized scale was motivated by convenience but it could be consequential. A user will probably care about some properties more than others and so weighting them all to be within $\pm x$% of the target is not necessarily the best adapted for that.

Alán Aspuru-Guzik communicated: I just also wanted to add that there is a fantastic connection between your work and our previous work. We worked on what is called *deep molecular dreaming*.[1] In it, we take advantage of the full representability of the SELFIES representation[2] and the deep dreaming method developed by Google Scientists. In deep dreaming, you can freeze the neural network once you train a neural net forward (*e.g.*, structure to property). Then you optimize over the set of inputs (structures) to match a particular output (property). This effectively does inverse design from property to structure and provides interesting interpretability. Exploring the connections of deep dreaming to your architecture would be a cool direction.

1 C. Shen, M. Krenn, S. Eppel and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.*, 2021, **2**, 03LT02, DOI: **10.1088/2632-2153/ac09d6**.
2 M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045024, DOI: **10.1088/2632-2153/aba947**.

Brett M. Savoie communicated in reply: Thanks for making this connection. It gave me an excuse to revisit that very creative work. The interpretability of this approach is quite a nice aspect. There are advantages to working with the properties directly in the LPM architecture, but you still face standard interpretability challenges.

Christian Kuttner communicated: If I understand correctly, the hypothesis is that providing a diverse set of properties during training might make the property-to-structure mapping unique and more reliable. What computational (or experimental) methods can be employed to verify the hypothesis that multiple

properties enhance the uniqueness and accuracy of the inverse mapping in LPMs?

Brett M. Savoie communicated in response: Great question. I think this could be answered with a relatively straightforward case study in which the reconstruction accuracy was monitored against the number of properties provided to the model. The reconstruction accuracy (*i.e.*, a direct measure of how singular the distribution becomes for a given property vector) should be monotonic in the number of properties according to this hypothesis. Now that I have spelled it out, we are definitely on the hook to supply this in our planned follow up.

Eltjo Mante communicated: Thank you for your paper (**https://doi.org/10.1039/d4fd00113c**), it was great to read. When you say that some properties share mutual information (that measuring the value of the 2nd property increases the accuracy/information of the readings of both properties), how do you strike the balance between choosing 10 properties with as much mutual information as possible (high exploitation) *vs.* properties with almost no overlap (high exploration of chemical space). The former approach seems to improve the accuracy of information at the expense of gaining new knowledge, whereas the latter seems to allow you to maximize new information learned, at the price of forgoing accuracy.

Brett M. Savoie communicated in reply: Thanks for pointing this out. In one of our masking case studies we see some evidence that there is mutual information between the properties owing to the accuracy improvement when we don't mask. More studies are definitely needed to understand the implications of this. We don't think there is a trade-off between including diverse properties *versus* related properties assuming they are available. Nevertheless, there might be practical constraints that lead to trade-offs (*e.g.*, the user can only afford to generate n properties for training). Put differently, there is nothing stopping the user from supplying an excess of properties if they are available. My instinct says that including redundant properties has little downside, apart from necessitating some training strategies like masking to avoid memorization behaviors. This is the same rationale behind including "constraints" during training that are largely redundant from an information perspective.

Filippo Bigi opened discussion of the paper by Venkat Kapil: When fine tuning foundational models, there seem to be two advantages in terms of computational power compared to training from scratch. One is that fewer DFT/reference calculations are needed, the other is that the model needs to be trained for a smaller number of epochs. Which one provides the most computational savings, in your experience?

Venkat Kapil replied: The answer depends on the system sizes, which determine both the training time and the time required to generate the training set. Most total energy and force calculations scale at least as $O(N^3)$, while the training time for one epoch using a message-passing-based graph neural network scales as $O(N)$, where N is the number of atoms in the system, assuming all structures have the same number of atoms. Additionally, the training time for a single epoch also scales as $O(M)$, with M representing the number of training configurations. I have mostly worked with

moderately sized systems and small training sets (<1000 structures), and in this regime, the cost of generating the training set is the computational bottleneck.

Roohollah Hafizi asked: In comparing the bespoke model for water molecules with the fine-tuned MACE foundation model for water molecules, which was faster to train?

Venkat Kapil answered: In this case, we found that training the fine-tuned model was faster for the fixed training set and the total number of epochs due to the presence of relatively fewer parameters in the pre-trained model. This difference in parameters comes from the use of 'RealAgnosticInteractionBlock' in the from-scratch models as opposed to the 'RealAgnosticResidualInteraction-Block'. The details on the interaction blocks can be found in ref. 1. However, it is to be noted that in practice, when fine-tuning, you may need a smaller volume of training data, and thus, the training time may be lower trivially due to the smaller size of the training set.

1 I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky and G. Csányi, The design space of E(3)-equivariant atom-centered interatomic potentials, *arXiv*, 2022, preprint, arXiv:2205.06643, DOI: **10.48550/arxiv.2205.06643**.

Roohollah Hafizi enquired: Training machine learning models on water systems is relatively straightforward due to water's high dielectric constant, which screens most interactions beyond 6 Å. Will the model that is used in your paper (**https://doi.org/10.1039/d4fd00107a**) also be effective for larger molecules with more significant long-range interactions?

Venkat Kapil responded: The largest molecules that we have tested are from the X23 dataset, and we can confirm (work in preparation) that we observe sub-kJ mol^{-1} performance with the same protocol as presented in this paper (**https://doi.org/10.1039/d4fd00107a**). We can't say anything about larger molecules before testing. Still, I am optimistic as the receptive field of MACE becomes fairly large after one round of message parsing (twice the typical cutoffs) and hopefully captures the interactions at hand.

Chris J. Pickard said: The result that pretraining the network parameters permits such effective fine-tuning on relatively small datasets is fascinating. It is reminiscent of the importance of generating "random sensible structures" for AIRSS driven structure prediction – the network's parameters somehow being "sensibly" chosen by the pretraining process. In the case of AIRSS, if the structures are made too sensible, areas of configuration space are excluded, and important structures can be missed. Is there the chance of something similar occurring in the case of fine-tuning models? For example, might pretraining on low pressure materials cause problems for fine-tuning models for chemically very different high-pressure phases.

Venkat Kapil replied: Chris, your parallel is spot on. I see using a pre-trained model as a good initial guess for the weights. Theoretically, the model's weights could get trapped in a local minimum near that of the pre-trained model.

However, in practice, the stochastic gradient descent algorithm can help the system move beyond local minima. That said, I agree with you – there may not be much benefit in initializing weights from a network trained on structures that are either very different or not sufficiently diverse.

Roohollah Hafizi requested: From previous experience, it seems that machine learning potentials often struggle to maintain bond lengths within "reasonable" limits. Could you clarify whether your model includes any additional terms, similar to those used in classical force fields, to help keep the water molecules connected?

Venkat Kapil responded: No, we haven't used any baseline to keep molecules intact. Based on my experience, machine learning potentials exhibit the artefact you mention when trained on inadequate or poor quality (noisy) data.

Janine George enquired: Can you anticipate how much of an impact the quality of the data used to train the foundation model will have on your fine-tuning approach?

Venkat Kapil answered: Short answer – a lot. Particularly if your machine learning interatomic potential is accurate enough to regress total energy and forces to an accuracy lower than the noise in the total energies and forces. As you can see, our density functional theory calculations use conservative plane wave cutoffs and hard pseudopotentials, ensuring the noise in our total energies and forces is down or lower than 1% of their standard deviations, as noted by our model's ability to reach such low errors. We were able to reach DFT accuracy at finite temperature with as few as 20 configurations.

On the other hand, the random phase expansion, based on hybrid-functional DFT, couldn't be converged fully with respect to the basis set, which required 75 configurations for stable *NVE* ensemble simulations due to the larger noise in total energies and forces.

Volker L. Deringer asked: In Fig. 1 of your paper (**https://doi.org/10.1039/d4fd00107a**), you show results for directly-trained and fine-tuned MACE models, as well as for directly-trained Behler–Parrinello neural-network (BPNN) potentials. Have you experimented with pre-training and fine-tuning BPNNs as well? It would be interesting to compare different ML potential fitting architectures side-by-side in this regard.

Venkat Kapil replied: We did not experiment with fine-tuning BPNNs. This was mainly because we concluded that the Behler–Parrinello neural network (BPNN) potentials for our selected symmetry functions may have reached saturation in training errors due to the limited expressivity of two- and three-body representations. However, we are aware of work by the groups of Berkelbach, Reichman, and Markland,[1] who have leveraged transfer learning for BPNNs for bulk simulations in the *NVT* ensemble at explicitly correlated theory levels.

1 M. S. Chen, J. Lee, H.-Z. Ye, T. C. Berkelbach, D. R. Reichman and T. E. Markland, Data-efficient machine learning potentials from transfer learning of periodic correlated

electronic structure methods: liquid water at AFQMC, CCSD, and CCSD(T) accuracy, *J. Chem. Theory Comput.*, 2023, **19**(14), 4510–4519, DOI: **10.1021/acs.jctc.2c01203**.

Liam Marsh questioned: In your slides, in the side-by-side comparison of fine-tuning *vs.* learning from scratch, the learning curve of the from-scratch model improves monotonically, but the fine-tuned one doesn't. For a couple of the points of this learning curve, it performs worse than the "previous best" point along said curve. Is this a possible artefact of the fine-tuning procedure, or is it just a coincidence?

Venkat Kapil answered: Thanks for the question. It is worth noting that the energy errors are already very small, and this regime, one of the from-scratch or fine-tuned models, could be made smaller by altering the relative weights of the energy and force loss. I personally wouldn't take this small deviation very seriously, and it is likely a random occurrence rather than something systematic.

Vishank Kumar enquired: What is the best way to sample the chemical space for fine tuning these foundational models? Do you suggest high pressure or high temperature *ab initio* molecular dynamics (AIMD) runs for sampling? Are there standard protocols/workflows for sampling and benchmarking these potentials when they are tuned before doing actual physics?

Venkat Kapil replied: That depends on the specific problem at hand, unfortunately. The sub-sampling problem is NP-hard,[1] meaning that only heuristics can provide solutions with satisfactory performance within a reasonable timeframe. In this case, we are focusing on a single temperature–pressure state point, so an *ab initio* molecular dynamics run at that state point suffices. However, we have experienced training potentials for the full-phase diagram, where we employed an active learning scheme to explore new temperatures and pressures.[2] More generally, it can be inferred that a training set taken from high-temperature simulations may capture more diverse local environments, potentially leading to a more generalizable model.[3] I am not aware of workflows for testing these, as fine-tuning is a very new field.

1 B. K. Natarajan, Sparse approximate solutions to linear systems, *SIAM J. Comput.*, 1995, **24**(2), 227–234, DOI: **10.1137/s0097539792240406**.
2 V. Kapil, C. Schran, A. Zen, J. Chen, C. J. Pickard and A. Michaelides, The first-principles phase diagram of monolayer nanoconfined water, *Nature*, 2022, **609**(7927), 512–516, DOI: **10.1038/s41586-022-05036-x**.
3 B. Monserrat, J. G. Brandenburg, E. A. Engel and B. Cheng, Liquid water contains the building blocks of diverse ice phases, *Nat. Commun.*, 2020, **11**(1), 5757, DOI: **10.1038/s41467-020-19606-y**.

Claudia Draxl asked: Molecular crystals often exhibit properties that are very different from those of the individual molecules. Also, the energy landscape of condensed molecular phases is typically very shallow, leading to the formation of different polymorphs, as discussed in your work. Is there any chance that we will be able to learn and predict the crystalline phases and their properties from the molecules themselves?

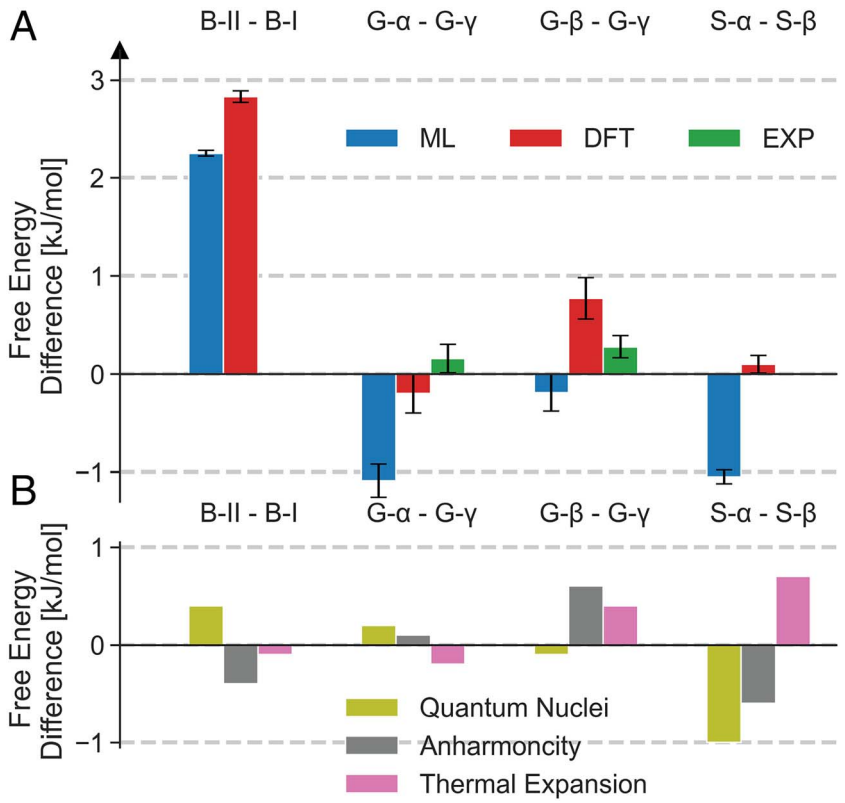

Fig. 1 (A) Path-integral (PI) Gibbs free energy differences between forms II and I of benzene (B-II and B-I); α-, β-, and γ-glycine (G-α, G-β, and G-γ); and α- and β-succinic acid (S-α and S-β) calculated using PBE0-MBD-based MLPs (blue) with the quantum thermodynamic integration (QTI) approach and corrected to the *ab initio* PBE0-MBD DFT level using free energy perturbation (red). Experimental data are shown in green. (B) Contributions of quantum nuclei (olive), anharmonicity (gray), and cell expansion and flexibility (pink) to the relative stabilities of the said polymorphs. These have been respectively obtained by comparing Gibbs free energy differences to estimates from a classical thermodynamic integration, a harmonic approximation, and a quantum thermodynamic integration using a fixed 0-K optimized cell. Reproduced from V. Kapil and E. A. Engel, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2111769119.[1]

Venkat Kapil responded: Indeed. We already tried to tackle this problem a few years ago using machine learning interatomic potentials based on high-dimensional artificial neural networks for molecular crystal polymorphs of benzene, glycine and succinic acid.[1] In that work, our accuracy was above 1 kJ mol^{-1}, and we needed to perform free energy perturbation to the density functional theory level to get an agreement with the experiment (see Fig. 1 here). We are currently using the MACE architecture to revisit this problem to predict polymorph stabilities, starting with the X23 dataset.

1 V. Kapil and E. A. Engel, A complete description of thermodynamic stabilities of molecular crystals, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**(6), e2111769119, DOI: **10.1073/pnas.2111769119**.

Alán Aspuru-Guzik enquired: Great talk! While fine-tuning these force fields, does it happen that your previously-trained model loses generalizability to other molecular systems? If so, is there a way to preserve this generalizability while fine tuning?

Venkat Kapil replied: Thanks, Alán! Yes, our 'naive' fine-tuning approach, which uses the final checkpoint of the pre-trained model to initialize weights, is susceptible to catastrophic forgetting. This is not an issue for this paper (**https://doi.org/10.1039/d4fd00107a**), as we are interested in a purpose-specific model trained by adding data relevant to the system we want to simulate. However, if the goal is to retain the generalizability of the pre-trained model, there are ways to do so, which revolve around limiting large changes in weight, *e.g.*, using a regularization that penalizes large changes in the weights with respect to those of the pre-trained model.[1] Another approach, currently a work in progress, is to have multiple readouts with a shared backbone, with one of them predicting the total energies and gradients of the training set. In contrast, the other readout predicts those of a subset of the pre-trained dataset. The loss function penalizes deviations for both the readouts. This approach is implemented in the MACE code.[2]

1 J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran and R. Hadsell, Overcoming catastrophic forgetting in neural networks, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**(13), 3521–3526, DOI: **10.1073/pnas.1611835114**.
2 **https://github.com/ACEsuit/mace**.

Shirui Wang asked: From your perspective, what role has the number of epochs played in your fine-tuning, especially regarding the fine-tuning of MACE models? What is the largest number of epochs you tried?

Venkat Kapil answered: We used 1000 epochs to fine-tune MACE-MP-0 to a purpose-specific model, ensuring the best density and potential energy performance at finite temperatures. This is the recommended strategy for developing a purpose-specific model, as one is not concerned about the model losing accuracy on the pretraining structures. However, in general, if a model is fine-tuned naively for too many epochs by simply using the pre-trained model checkpoint to initialize model weights, it may 'forget' previously learned information. Therefore, a different fine-tuning strategy should be adopted if the goal is to improve the model while preserving its accuracy on the pretraining data.

Shirui Wang enquired: Did you use any empirical methods to check whether your model is overfitting, particularly when using a large number of epochs?

Venkat Kapil responded: We studied the validation error as a function of training epochs and didn't observe an increase in the error.

Christian Kuttner asked: The method demonstrated good agreement with experimental data for hexagonal ice at the random phase approximation level. How might this protocol be adapted or extended to improve the modeling of other molecular crystals or different types of polymorphs?

Venkat Kapil replied: Many thanks for this question. To enable sublimation enthalpy predictions for other molecular crystals, larger and more complex than ice, we are exploring whether it might be sufficient to train on clusters extracted from the bulk. Clusters could be more suitable for high-level electronic structure theory calculations than periodic systems.

Branko Ruscic said: Let me point out several things. Firstly, the standard for the expression of uncertainty in thermochemistry – universally followed by thermochemists and by all thermochemical tabulations worth their salt – is at the level of 95% confidence intervals, rather than at the level of standard deviations or root mean square errors, or, even worse, mean absolute deviations (see discussion in ref. 1 and 2). The intention of the accepted standard in thermochemistry is that the true answer is not outside the quoted error bar more than once in every twenty cases. It is important for everybody to follow this standard, or else intercomparison of data and their perceived accuracy becomes misleading. In order to conform to the thermochemical standard, your quoted accuracy for predicted sublimation enthalpies, which is apparently based on the root mean square errors, should be augmented approximately by a factor of 2. A second point is that the thermochemical sublimation/vaporization enthalpies always refer to the target chemical species in the ideal gas state, rather than real gas vapor, where the latter includes not only the monomer, but also the dimer, trimer, *etc.* Thus, there is a difference between vaporization/sublimation enthalpies that may have been obtained directly from experimental determinations of equilibrium pressures, by calorimetry, by weight loss, *etc.*, and those that are properly corrected in order to follow the thermochemical definition: the difference between the two is the real-to-ideal correction, typically obtained by an analysis of virial coefficients or some other equation of state (see ref. 3). The real-to-ideal corrections are frequently of the order of a significant fraction of a kJ mol^{-1}, but in some cases can be as high as several tens of kJ mol^{-1} (*e.g.* acids, such as formic, acetic, HF, *etc.*). Admittedly, sublimation means that the reactant is in the solid phase, implying that in most cases the equilibrium pressure at room temperature is presumably rather low, preventing the formation of dimers and higher oligomers, and consequently frequently making the real-to-ideal correction negligible at 298.15 K, although even for solids this needs to be carefully checked. One should also state here that computational vaporization/sublimation enthalpies, as long as they are based on the difference between computed thermochemical properties of the ideal gas and condensed phase, intrinsically follow the thermochemical definition, but their benchmarking with experimental values becomes rather complicated, since one has to dig deeply into the provenance of each experimental point in order to distinguish between corrected and uncorrected benchmarks. The third point is that in ML, generally speaking, the data in the learning set will tend to contain both statistical noise and outright outliers. Thus, if the ML model seems to perform better than the average uncertainty of the learning set, it is rather likely that the learning set was overfitted, and that the resulting ML model faithfully reproduces the entirety of the learning set, warts and all, faithfully reproducing the outliers and other errors of the learning set.

1 B. Ruscic, Uncertainty quantification in thermochemistry, benchmarking electronic structure computations, and Active Thermochemical Tables, *Int. J. Quantum Chem.*, 2014, **114**, 1097–1101, DOI: **10.1002/qua.24605**.
2 B. Ruscic and D. H. Bross, Thermochemistry, *Comput.-Aided Chem. Eng.*, 2019, **45**, 3–114, DOI: **10.1016/B978-0-444-64087-1.00001-2**.
3 B. Ruscic, Active Thermochemical Tables: water and water dimer, *J. Phys. Chem. A*, 2013, **117**, 11940–11953, DOI: **10.1021/jp403197t**.

Venkat Kapil responded: Thanks for these comments. Next time we do a sublimation enthalpy prediction, we will be sure to report 95% confidence intervals and clearly report if the sublimation is based on an ideal gas reference. Regarding your third point, we have ensured that the noise in our training data is very low so that we don't have the issue of overfitting to noise. I agree that it's an important point.

Christian Kuttner communicated: The approach uses only a few tens of training structures to achieve high accuracy. What are the key factors that enable such a low number of training structures to provide accurate potential energy surfaces and finite-temperature sampling?

Venkat Kapil communicated in reply: The first factor is the use of the atomic cluster expansion[1] basis in the MACE architecture, which provides a smooth and equivariant basis set for learning the potential. The second factor is the use of message passing, which enables the construction of high body-order representations for regressing total energies and forces. These attributes make MACE more accurate and better at extrapolation than previous non-equivariant and truncated body-order machine learning interatomic potentials, leading to stable finite-temperature simulations even when trained on small datasets.

Our approach improves upon MACE MLIPs trained from scratch by fine-tuning the pre-trained MACE-MP-0 model. Simply put, the pre-trained model initializes weights that are already much 'closer' to the optimal solution for the new problem, thereby requiring less data to reach optimal weights. This can be observed from the initial loss function in 'from scratch' *versus* 'fine-tuned' training, which is significantly smaller for the latter. Another factor that enables us to learn from a small volume of data is the use of high-quality, low-noise data. This is achieved by being overly conservative with density functional theory (DFT) parameters, such as the plane wave cutoff, *k*-point grid, and pseudopotentials.

1 R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, *Phys. Rev. B*, 2019, **99**(1), 014104, DOI: **10.1103/physrevb.99.014104**.

Jennie Martin opened discussion of the paper by Takuya Taniguchi: Regarding selection of teacher and student models, can successful knowledge distillation be achieved with any well-performing teacher model and relevant student model or are there other considerations impacting whether or not a teacher and student are a good pairing for knowledge distillation to succeed?

Takuya Taniguchi replied: I'm not entirely certain, but in my opinion, it depends on the structure of the student model and the domain of the data you want to apply it to. For example, if you try to apply the knowledge from a highly accurate, complex teacher model to an overly simple student model, it might work

well if you limit the data domain very narrowly. However, if the data domain is too broad for the student model to fully learn, the knowledge transfer might not function as effectively. There isn't necessarily a perfect combination; rather, I believe the success of knowledge transfer is determined by the balance between the teacher model, the student model, the breadth of the data domain you want to transfer, and the amount of training data for the transfer.

Shubham Vishnoi asked: How are organic crystals defined within the training dataset used in this study? What criteria are used to screen these structures as organic? Does the training dataset include structures like CCl_4 or amino acids with sulfur, or are such structures excluded?

Takuya Taniguchi responded: Based on the MPtrj dataset containing 1.58 million structures, organic crystals comprise 1.8% of the total, specifically 28 402 data points. Organic crystals are defined as structures composed exclusively of H, C, N, O, P, S, F, Cl, Br, and I. While compounds like CCl_4 or sulfur-containing amino acids would technically be classified as organic under these criteria since they only contain elements from the allowed list, I don't believe these specific compounds were present in the dataset, though I haven't verified every single structure.

Shubham Vishnoi enquired: Does the training dataset used in this study include structures labelled as 'theoretical' from the Materials Project? Additionally, how does the dataset define organic and complex crystals, and what impact do these classifications have on the neural network potentials training process?

Takuya Taniguchi answered: The dataset used in this study is theoretical, as it includes not only stable structures but also trajectory data. Regarding the classification of crystals in the dataset, the paper (**https://doi.org/10.1039/d4fd00090k**) defines organic crystals as structures composed solely of H, C, N, O, P, S, F, Cl, Br, and I. This is a standard criterion for organic crystals. Complex crystals, being neither organic nor inorganic, were defined according to the conditions stated in the paper. The majority of the dataset consisted of inorganic crystals, with organic crystals accounting for only about 2% (and no complex crystals). While it's difficult to clearly determine how the differences in definitions contributed to the learning process, this study investigated how the neural network potential (NNP) could be improved using the soft targets of the teacher model, even with limited organic crystal data.

Janine George questioned: I would assume that the discrepancy in the volume prediction for organic crystals from Crystal Hamiltonian Graph neural Network (CHGNet) is partly related to the lack of dispersion correction of the data underlying the CHGNet model. Have you considered correcting the CHGNet results with an empirical dispersion correction?

Takuya Taniguchi responded: Yes, the discrepancy in CHGNet's volume predictions for organic crystals is likely due in part to the lack of dispersion correction in the original training data. This study does not directly apply empirical dispersion corrections, but instead adopts a knowledge distillation

method to indirectly learn the effects of dispersion forces. By transferring knowledge from preferred potential (PFP), a model that considers dispersion forces, the prediction accuracy of CHGNet is improved. The application of empirical dispersion corrections could be a promising approach for future research, and combining it with knowledge distillation might lead to further improvements.

Yuchen Lou asked: The student model learns from both soft and hard targets, but are you worried that the student model becomes too dependent on the teacher model and simply reproduces what the teacher model is predicting?

Takuya Taniguchi answered: I agree with your comment. The learning results are likely to be highly dependent on the performance of the teacher model. Since we performed additional training on the pretrained student model to minimize the loss, including the difference with the soft target, the student model will try to approach the teacher model, at least within the range of the training data. However, since the student model and the teacher model have different model architectures, complete imitation is difficult. Rather, it can be considered that the knowledge of the teacher model is being learned in a form suitable for the structure of the student model.

Filippo Bigi enquired: Foundational models are relatively slow compared to other models due to their size and need to learn from large and diverse datasets. Do you think it would be possible to perform knowledge distillation from a large pre-trained model to a faster model that is specialized for a specific application?

Takuya Taniguchi replied: Knowledge distillation from large pre-trained models to faster, application-specific models is possible. This approach offers advantages such as improved computational efficiency, reduced resource requirements, and enhanced performance in specific tasks, while also enabling easier execution on smaller devices. However, it comes with challenges like reduced versatility, potential loss of some knowledge, and the importance of appropriate task selection. Overall, I think that this method is a promising approach to balance computational efficiency and performance improvement in specific tasks, but it requires careful model design and task selection.

Christian Kuttner communicated: Knowledge distillation from a teacher model has been demonstrated to enhance the accuracy of NNPs for organic molecular crystals. What specific aspects of the teacher model's soft targets contribute most significantly to the improved accuracy, and how can these aspects be systematically optimized for different types of organic molecular crystals?

Takuya Taniguchi communicated in reply: The main factor contributing to the improved accuracy of NNPs for organic molecular crystals through the teacher model's soft targets is likely the sensitivity to intermolecular interactions captured by the teacher model. In particular, the representation of non-covalent interactions such as hydrogen bonding, van der Waals forces, and dispersion forces is crucial. These interactions significantly influence the structure and

properties of organic molecular crystals, but NNPs trained primarily on inorganic crystal data may not adequately capture them. Regarding systematic optimization methods, I don't have a clear answer, but if the model's expressive power is sufficient, increasing the quantity and diversity of data could be considered. In other cases, model merging, which integrates multiple models into a single model, might be an option. For model merging, individual models can be trained on small datasets, and then integrated to create a model with high generalization performance.

Christian Kuttner communicated: Your study found that increasing the ratio of hard target loss in knowledge distillation leads to overfitting and reduced efficiency in knowledge transfer. Could there be alternative methods or modifications in the training process that mitigate the risk of overfitting while maintaining the benefits of knowledge transfer for NNPs applied to organic molecular crystals?

Takuya Taniguchi communicated in response: To prevent overfitting in neural networks, techniques such as regularization and dropout are commonly used, and these are expected to be effective in controlling overfitting in NNP training as well. L1 and L2 regularization control the behavior of neural network weights by adding penalty terms to the loss function. Dropout randomly deactivates a portion of neurons during training, preventing the model from overfitting to specific features or noise. In addition to these methods, if it's possible to add more data, increasing the amount and diversity of data might improve generalization performance. However, the effectiveness of these techniques depends on the model and dataset, so it's impossible to know without testing whether they will successfully prevent overfitting and improve generalization performance.

Kevion K. Darmawan opened discussion of the paper by Veronika Juraskova: Magnesium (Mg) is an essential co-factor for many biochemical reactions in the human body. For instance, it is pivotal for RNA polymerase enzymatic activity and ATP hydrolysis, both of which occur at physiological temperature. Is there a specific reason for running the simulation at room temperature instead? Would the umbrella sampling calculation be affected if the temperature were changed? Additionally, is there a particular reason why the simulations were performed using the *NVT* ensemble instead of the *NPT* ensemble?

Veronika Juraskova answered: In this work (**https://doi.org/10.1039/d4fd00140k**), we studied the $[Mg(H_2O)_6]^{2+}$ complex as a model system for the ligand exchange in water solution. As we were interested solely in the Mg^{2+} in solution and not in its physiological behaviour, we decided to simulate it at room temperature. Moreover, the experimental data for water exchange are reported for room temperature as well. Regarding your second question, the rate constant of the ligand exchange, and corresponding barrier, are temperature-dependent quantities, therefore there is an impact of the temperature on the free energy profile. The extent of this impact can be directly extracted from the experimental data in ref. 1, which reports the fit of the ^{17}O NMR reaction rates as a function of the temperature and pressure. As there is only a minor

impact of the pressure on the free energy barrier in an aqueous solution, we selected the *NVT* ensemble for its simplicity.

1 A. Bleuzen, P.-A. Pittet, L. Helm and A. E. Merbach, Water exchange on magnesium(II) in aqueous solution: a variable temperature and pressure ^{17}O NMR study, *Magn. Reson. Chem.*, 1997, **35**, 765–773.

Philip C. Biggin asked: You have shown very nice validation results, but then performance of the potential of mean force (PMF) profiles in Fig. 4(a) in your paper (https://doi.org/10.1039/d4fd00140k) shows that the peak is about 2 kcal mol^{-1} off the expected/experimental value and Fig. S13 in the supplementary information shows that a repeat does not change this. Do you have any sense of how you could improve it to get closer to the experimental value? Is there some aspect you suspect you are not sampling?

Following on from that, what kind of error would you be satisfied would be useful for biological simulation going forward? Finally – how easy would it be to do Mg with ATP given that is the most relevant biological chelator for Mg^{2+} and has very different environment than just water of solvation?

Veronika Juraskova replied: There could be several sources of error in the reaction barrier in Fig. 4(a) in the paper (https://doi.org/10.1039/d4fd00140k), such as the use of ωB97X-D3BJ/def2-TZVP level of theory, missing nuclear quantum effects, and selected sampling technique. Based on work by Schwierz *et al.*,[1] the cause of the error in the barrier might come from the fact that we compute the 1D PMF along the distance between Mg^{2+} and one water molecule as a collective variable. The results they reported suggest that the water exchange around the Mg^{2+} ion is a more complex process, involving the motion of several water molecules. The way to improve the accuracy could thus be to perform the free energy computations using a distance to $2H_2O$ as a collective variable or to use transition path sampling to sample the water exchange process. Concerning the errors, the reported experimental rate constants of the exchange reaction at room temperature vary from 5.3×10^5 s^{-1} ($\Delta H^{\ddagger} = 10.2$ kcal mol^{-1}, $\Delta S^{\ddagger} = 2$ cal K^{-1} mol^{-1}) to $6.7 \pm 0.2 \times 10^5$ s^{-1} ($\Delta H^{\ddagger} = 11.7 \pm 0.2$ kcal mol^{-1}, $\Delta S^{\ddagger} = 31.1 \pm 2.2$ cal K^{-1} mol^{-1}, $\Delta V^{\ddagger} = 6.7 \pm 0.7$ cm^3 mol^{-1}),[2,3] corresponding to exchange barriers of 9.6 kcal mol^{-1} and 9.7 ± 0.3 kcal mol^{-1}, respectively. The error within 1 kcal mol^{-1} from the reported values will be considered sufficient. Finally, modelling of Mg^{2+} complex with ATP is more challenging than modelling $[Mg(H_2O)_6]^{2+}$ for several reasons. Firstly, the system would have a significantly larger number of chemical elements, *i.e.*, 3 in the $[Mg(H_2O)_6]^{2+}$ complex *vs.* 6 in the ATP–Mg complex in solution, which would increase the number of basis functions and structures for the training, resulting in a large memory requirement and training cost. Furthermore, ATP is a larger and more flexible molecule with many stable conformations and several bonding sites for Mg^{2+} and H_2O. All these structures would need to be considered in the training set to ensure a reliable description of the stability of resulting complexes, requiring larger data sets and training strategies able to sample these conformations. Finally, ATP is not stable in the gas phase computation at GGA and hybrid GGA level of theory, the modelling strategy presented in this work would thus need to be adapted to avoid

introducing structures too different from the target potential energy surface (PES).

1 N. Schwierz, Kinetic pathways of water exchange in the first hydration shell of magnesium, *J. Chem. Phys.*, 2020, **152**, 224106.
2 J. Neely and R. Connick, Rate of water exchange from hydrated magnesium ion, *J. Am. Chem. Soc.*, 1970, **92**, 3476–3478.
3 A. Bleuzen, P.-A. Pittet, L. Helm and A. E. Merbach, Water exchange on magnesium(II) in aqueous solution: a variable temperature and pressure ^{17}O NMR study, *Magn. Reson. Chem.*, 1997, **35**, 765–773.

Matthew Dyer enquired: I was interested in the human-in-the-loop aspect of the active learning procedure that you applied, in which domain experts are used to see where the preliminary model was failing and to then choose which data to add to improve that. Do you see this as a strength of the approach, or would you like to automate this in the future?

Veronika Juraskova responded: At the current stage, we see the need for human intervention in the training cycle as a weakness. The training strategy we are aiming at should be robust and fully automated to avoid as many manual checks as possible. However, the final ML potentials will still need to be properly validated to ensure accurate performance across the required chemical and conformational space.

Heather J. Kulik asked: I'm interested to know more about the method choices for the Mg system *vs.* Pd system in generating training data. I'm interested to know what the rationale was for switching the functional between the two systems. Because I'm interested to know when choosing a DFT functional, how do you know your potential isn't going to go into a space where the functional performs poorly? What is the reference used for the Pd benchmark for selecting the functional?

Veronika Juraskova answered: To identify the most suitable level of theory for the ligand exchange, we performed a benchmark for the dissociation energy of a similar Pd-containing complex as reported here, namely $[Pd(en)(Py)_2]^{2+}$. For this complex, we generated geometries using the relax scan along the pyridine dissociation for Pd–N distance in a range from 1.5 to 3.25 Å at the PBE0-D3BJ/def2-SVP level of theory in the gas phase and recomputed energies for these structures at DLPNO-CCSD(T)/def2-TZVPP. We then benchmarked several DFT methods against this reference. Among them, the best performance was obtained for ωB97X-D3BJ/def2-TZVP, with comparable performance for the TPSS0-D3BJ/def2-TZVP level of theory. This result agrees with the previously observed good performance of ωB97X-D3BJ on main-group thermochemistry, kinetics, and noncovalent interactions.[1] We, therefore, selected ωB97X-D3BJ/def2-TZVP as a ground truth DFT method for the Mg^{2+} complex. However, as this method has a significant computational cost for large systems, we switched to TPSS0-D3BJ/def2-TZVP as a reference for the Pd^{2+} complex with MeCN solvent, which contains more heavy atoms than H_2O.

1 A. Najibi and L. Goerigk, The nonlocal kernel in van der Waals density functionals as an additive correction: an extensive analysis with special emphasis on the B97M-V and ωB97M-V approaches, *J. Chem. Theory Comput.*, 2018, **14**, 5725–5738.

Filip T. Szczypiński enquired: You have applied machine learning force fields to validate the experimentally determined solvent exchange mechanisms: one associative (for acetonitrile exchange at a palladium centre) and one dissociative (for water exchange at a magnesium centre). Have you been able to identify a preference for a given mechanism – *e.g.*, by evaluating both of them and comparing reaction barriers or by investigating a 2D reaction plot with the two mechanisms along the edges – using your methods *a priori*?

Veronika Juraskova replied: Thank you for a very good point. In this particular work (**https://doi.org/10.1039/d4fd00140k**), we evaluated the barriers only for the experimentally proven mechanisms and not for the complementary ones. However, modelling of the other mechanisms is indeed possible, and it would be interesting to see the differences in the barrier of both processes.

Filip T. Szczypiński asked: How difficult is it re-train your models to other metal-containing systems? Have you tried to apply transfer learning from the magnesium–water system to the palladium–acetonitrile system to accelerate the process without having to re-train from scratch? Alternatively, did you explore adapting to related metals within the same group where the electronic differences might be less profound: palladium to platinum or magnesium to calcium?

Veronika Juraskova answered: The strategy should be transferable to other similar metal-containing systems. We have not tried the transfer learning strategy for different metal ions, but this is planned for future work.

Ian Fairlamb said: Regarding the modelling with $[Pd(CH_3CN)_4]^{2+}$: practical applications, particularly in catalysis, involving this dicationic Pd^{2+} species would be affected by halide anions. Is there the potential to consider the impact on Pd^{2+} speciation by modelling in certain quantities of chloride, for example? I anticipate chloride having a high affinity for the Pd^{2+}, leading to speciation. If this could be modelled with the method described in your paper (**https://doi.org/10.1039/d4fd00140k**), then a stronger link to applied catalysis could (potentially) be made. There is further the interesting question of $[Pd–Cl–Pd]^{x+}$ species, in addition to higher order salt clusters. It is also interesting to consider the interaction of Pd^{2+} species with other metal halide salts like CuCl and AgCl, linked to Wacker oxidation chemistry.

Halide redistribution has been seen previously, see ref. 1. The implications for catalysis are significant.

1 K. L. Bray, I. J. S. Fairlamb and G. C. Lloyd-Jones, *Chem. Commun.*, 2001, 187–188, DOI: **10.1039/B009356O**.

Veronika Juraskova responded: Thank you for the reference and for highlighting the importance of the Pd^{2+} interactions and $[Pd–Cl–Pd]^{x+}$ species. While we focused solely on the complexation of Pd^{2+} cation by solvent molecules, an extension of the training data set towards modelling of the Pd–Cl complexes is

indeed possible, including the study of the impact of different Cl^- concentrations and Pd^{2+} speciation.

Venkat Kapil said: Quantum nuclear effects (NQEs) influence the barriers to aqueous chemical reactions, particularly those involving the possibility of water dissociation. It would be interesting to calculate the differences due to NQEs and whether there are experiments on water isotopomers for direct comparison.

Veronika Juraskova responded: Thank you for the suggestion. We did not try to include the NQE description in this work, but I agree that it would be interesting to see its impact, especially on the properties of Mg^{2+} in an aqueous solution.

Adarsh V. Kalikadienasked: You mentioned that you are aiming for full automation of your workflow. Could you comment on how much human intervention is needed in the process and where? For example, what parameters change when different metals or ions are taken into account? Are you aiming to address this by implementing a general set of rules?

Veronika Juraskova answered: The final training data set will be reliable in the simulations only if it sufficiently covers the chemical and conformation space of the problem we want to study. To ensure this, the dynamics in the active learning loop must be able to sample structures along the chemical process, including the reactants, transition states and products, as well as the path which connects them. While we always have some preliminary insights helping to design the active learning, chemistry in solution is often more complex than expected, containing many competing reaction channels which can be accessible during the sampling. When this situation occurs, it is necessary to properly validate the machine-learning potentials (MLP) in the unexpected regions and possibly adapt the active learning (AL) to account for them. This is the part where human intervention is typically needed. We are planning to implement strategies which will help to handle these situations, such as enhanced sampling and on-the-fly uncertainty estimations.

Barnabas A. Franklin communicated: Given the successful applications to both Mg and Pd, where do you see this heading in the future and what do you think the challenges will be if scaling this methodology to larger or more complex systems? Do you foresee any issues with data availability when investigating less well documented metal complexes?

Veronika Juraskova communicated in reply: In my perspective, the key challenge in the applicability of machine learning potentials in general lies in the availability of suitable datasets for training. As we generate the datasets from scratch, we can bypass the issues with data availability. However, larger and more complex systems cover larger chemical space, demanding a larger number of data points, which will push the present active learning strategy to its limits. Furthermore, once we start to model less-documented metal complexes, we need to be especially careful with the ground truth electronic structure and MLP validation, as the experimental reference will become scarce. In this case, we will

need to rely on the accuracy and predictive power of the underlying *ab initio* models.

Christian Kuttner communicated: The research successfully models the structural and dynamic properties of metal ions such as Mg^{2+} and Pd^{2+} in different solvents. What are the key challenges in extending this methodology to more complex systems, including those with multiple metal ions or mixed solvent environments?

Veronika Juraskova communicated in response: The extension of the presented work (**https://doi.org/10.1039/d4fd00140k**) to a mixture of solvents would be straightforward by adding clusters containing the mixture of the solvents. More challenging would be the modelling of complexes with large and flexible ligands, which might require additional strategies to generate the data sets encompassing sufficiently the conformational space, as was discussed in the case of ATP. Last but not least, the MLP reproduce the underlying ground truth electronic structure, the substantial challenge thus lies in the selection of the level of theory which provides a reliable description of the electronic structure of metal complexes and their interactions.

Varinia Bernales communicated: During your presentation, you mentioned that you aim to apply this methodology and the associated tools to model metalloenzymes, as well as iron- and manganese-containing metal complexes. Given that you used different levels of theory to represent Mg and Pd complexes based on the observed performance of the selected functionals for these systems, what would be your strategy for training the MLPs in cases where DFT fails due to the rise of static correlation, such as in manganese- or iron-based complexes? Additionally, how could we achieve a more transferable approach when multiple systems are required within the same simulation?

Veronika Juraskova communicated in reply: Thank you for this interesting question. We are aware of the complicated electronic structure of transition metal (TM) complexes and the possible sensitivity of the spin state to the coordination environment. We might still be able to use the DFT, MP2 or DLPNO-CCSD(T) methods in the case where the multireference character of the selected complexes is low; however, this will require careful analysis of the stable electronic states of the TM in active sites before we even start to think about the MLP training. The application of multireference methods to large metal complexes, including complexes in solution, is an ongoing research question and we are excited to see and try the latest developments in more challenging cases.

Wenhao Sun addressed Christopher M. Collins and Chris J. Pickard: Even though FUSE and Hot AIRSS tackled these very complicated structures, I think the most complicated structures for crystal structure prediction here are actually elemental Mn and elemental boron. The unit cell of elemental Mn is 29 atoms, and is famously complicated. I anticipate that this complexity arises because of the anti-ferromagnetism, meaning that the different Mn atoms in the unit cell should perhaps not be treated by a single interatomic potential, but rather by different potentials for each spin-up or spin-down species. Similarly, in boron it

has been argued that the complexity of the crystal structure comes from the internal charge separation of boron into B^{3-} and B^{5+}; in essence boron makes an ionic solid with itself. My point is that if the ML interatomic potentials do not distinguish between these two species, which are nominally the same element type, I think the accelerated structure generation might push the system down into unphysical 'energy wells'. More generally speaking, what are the consequences if there are salient physics governing crystal structure selection that are not accounted for in the training crystal structures for MLIP training; or if the architecture of the MLIP does not account for important interactions (long-range electrostatic, magnetic, *etc.*)?

Christopher M. Collins responded: I think generally, speaking, when training MLIPs, because there are a good number of these interactions indeed captured (even if for the wrong reasons!), the quality of these will certainly improve with time as more complex electronic structure calculations are used as the basis for training data, for example, where we are starting to see MLIP models which are specifically trained on spin polarised calculations.

With regards to the consequences of a MLIP, which is not trained correctly, there are two main potential problems:

(i) That the energy rankings can be wrong, and this can potentially even be very subtle differences in structure, for example, with distortions caused purely by magnetic structure.

(ii) The forces end up being learnt incorrectly, which could result in what should be stable minima not being local minima at all on a potential energy landscape, which could in theory prevent a ground state from being located.

I think in practice so far though, neither of these problems seem to be appearing, especially as with Chris Pickard's paper (**https://doi.org/10.1039/d4fd00134f**), where using their interatomic potentials, they can locate the correct minima. Equally with the Mn system in the FUSE paper, this does not appear to be a problem, if we take the experimental structure of Mn and optimise it with the potentials used within the paper and the corresponding (non-spin polarised) density functional theory calculation, the experimental structure remains intact, and lower in energy than the structures we obtained in this benchmark.

Chris J. Pickard replied: I agree that the boron example is chemically complex – and as such is a challenging test system for both structure search and MLIPs. In the original EDDP paper[1] it was shown that potentials fitted to training data containing just 8 boron atoms in the unit cell could recover both α-boron and the more complex β-boron, which consists of $B_{12}^{\delta-}$ icosahedra and interstitial $B_2^{\delta+}$ dimers. More recently, in unpublished work, I have seen that the same result can be obtained using training data with just 4 atoms in a unit cell. This is a remarkable result. These small unit cells cannot contain any intact B icosahedra. EDDPs are trained on a large number of highly structurally diverse, but small, configurations. Clearly this is sufficient to learn the favoured environments that can relax to low energy structures in larger unit cells. I have also tested the approach for elemental Mn, and EDDPs generated from 4 atom unit cells enable the rediscovery of the complex known phases using AIRSS. You are right though that there are limits to this – if the structural environment is identical between

sites with different magnetic ordering there is no way for a simple model to distinguish between them. The development of machine learning models that incorporate magnetic and electronic degrees of freedom is an active area of research. Similarly, the detailed treatment of long ranged interactions. It should be noted though that it is observed that truncated potentials work well in condensed phases, where the long-range interactions are either not dominant, or screened. This is not necessarily the case in more inhomogeneous systems.

1 C. J. Pickard, *Phys. Rev. B*, 2022, **106**(1), 014102.

Tim Bechtel addressed Christopher M. Collins: Is there a trade-off between biasing a dataset on diversity of different structures, *versus* having low formation energies? Does data diversity help with finding new, interesting chemistries/structures, at the cost of higher energy?

Christopher M. Collins answered: Currently, I think that this is the case, certainly in the case of crystal structure generation with ML models. From my experience at trying them, they are getting towards the position where they can generate sensible crystal structures, but the energy predictions are way off. In the context of my work of doing crystal structure prediction using heuristic crystal structure prediction algorithms, this then has the consequence of leading a searching algorithm into the wrong areas of the potential energy landscape. I think that in the future, I would like to see more of an emphasis on first of all being able to correctly predict/calculate energetics, then either use separate algorithms or models to explore the arrangement of atoms.

Claudia Draxl addressed Christopher M. Collins and Chris J. Pickard: High-throughput screening of stable crystal structures is a very popular field. Most of these calculations are performed with semi-local density functionals such as PBE. The results are typically rather quite robust. However, in complex structures with large unit cells, subtle effects may play a crucial role. For example, charge localization around an impurity can change bond lengths and thus the overall bonding situation. In such cases, hybrid functionals may give a different result. Can you estimate, what fraction of predicted structures cannot be trusted?

Christopher M. Collins responded: I would think that at the moment, the honest answer, is that it is difficult to assess this. In practice (apologies if this has already been done!), it would be to investigate large databases such as the materials project, and make these types of comparisons with experimental structures, perhaps collecting statistics on the numbers of areas and/or specific chemistries where the calculated structure does not agree with an experimental one. Of course, this would not necessarily provide information as to why there is a lack of agreement, as there could be any number of reasons, for example: some kind of temperature dependent effect on the structure, errors within the functional, or even experimental error in the measured structure. So I think at the moment, it would be difficult to put a number on this, but certainly in my area (ionic solids) it appears that the results for the vast majority of structures are reasonable.

Chris J. Pickard replied: I would approach all results with scepticism, given the approximations that density functional theory entail. The key question is whether any predictions are sufficiently robust to justify serious consideration. Much of the pioneering work in structure prediction was on the transformation of materials under pressure. There was an excellent success rate for these early predictions, but they relied on it being a rather forgiving problem. If a discovered phase is rather low in enthalpy, and somewhat denser than competing phases, the $+pV$ term in the enthalpy makes it very likely that it would become thermodynamically the most stable at some elevated pressure. Errors in the relative enthalpy of phases lead to errors in the predicted transition pressure. A prediction of a new phase appearing at 100 GPa would be considered successful even if the experimentalist had to reach slightly higher pressures to find it. The situation is much more challenging at ambient conditions, where a predicted phase only becoming apparent at 1 GPa would lead to a failed prediction. Density functional theory predictions are particularly unreliable when comparing phases of very different densities – so I would be particularly cautious about predictions of, for example, very open structures made with the PBE functional, which is known to favour low density phases.

Nicholas David addressed Chris J. Pickard and Christopher M. Collins: There are some monatomic crystal structures that contain complex physics which we can accurately predict, possibly for the wrong reasons, but others that we cannot. Which areas of chemical space and structure space does crystal structure prediction perform well and in which areas does it perform poorly?

Chris J. Pickard answered: We strongly depend on DFT for crystal structure prediction, because of its reasonable balance between accuracy and computational efficiency. MLIP accelerations still depend on DFT datasets for training. Beyond poor performance due to systems being excessively complex, where the "exponential wall" is encountered, any system for which DFT struggles will be difficult to perform structure prediction on, although DFT+U allows strongly correlated, and magnetic, systems to be treated to some extent.

Christopher M. Collins replied: I think that there are two parts which can define how effectively we can perform crystal structure prediction within a chemical space:

(i) The accuracy of the calculations which are performed, as even when we extend into using machine learnt interatomic potentials, we will not be able to exceed the accuracy of the calculation model used to train the potential.

In most cases for inorganic solids (I am not qualified to comment on organic structure prediction!), the main calculator is with an implementation of density functional theory (DFT) using PBE functionals. From my experience, this then performs well for strongly ionic systems, and the poorest performance is typically for systems which contain significant levels of covalent bonding. In situations where typical DFT setups are insufficient to describe the energy minima effectively, we can move to higher levels of theory, although this then comes with a significantly increased cost in terms of compute time.

(ii) Relating to the nature of potential energy landscapes: generally, the systems which are the most difficult to work with are those containing rigid

polyhedra which are readily rotated, for example, B^{3+} and Si^{4+} in trigonal planar or tetrahedral co-ordinations. These tend to be difficult as the energy landscapes often are relatively flat with many shallow energy minima, this is as a result of having many choices of orientation for these units, which are effectively degenerate. I would say that currently challenges caused by many shallow energy minima present the most significant challenge to crystal structure prediction.

Tim Bechtel addressed Christopher M. Collins: Can you quantify what accuracy you need the model to be, with respect to different derived properties, *e.g.* stable MD, convex hulls, phonon dispersions? *I.e.* do we know when to stop optimizing our model's mean absolute error (MAE)/root mean square error (RMSE) if we are after a specific property?

Christopher M. Collins answered: As a guideline, for energies, we would like to see errors within the range of ~10 meV per atom, this is at least the guidelines that we would like to get to. However, working towards just a target MAE/RMSE doesn't tell the whole story, especially in the context of calculating convex hulls. For example, it is entirely possible for a potential to have a very small MAE/RMSE, but, if one key reference compound has a wildly incorrect energy (say 5 eV per atom out, this is a large number, but we have seen it occur), then this will have the knock on effect of all of the convex hull energies that you calculate being very wrong. So I think when training ML models, it is also very important to look at the resulting distributions and look at outlying results, in addition to only considering training towards specific MAE/RMSE values.

Vishank Kumar addressed Veronika Juraskova, Takuya Taniguchi and Venkat Kapil: Did you train these models with AMD GPU's as compared to NVIDIA and if yes, did you notice any issue of dependency or speed?

Veronika Juraskova responded: We trained MACE models only with NVIDIA GPU cards.

Takuya Taniguchi answered: No, we have used only NVIDIA GPU (RTX 6000 Ada).

Venkat Kapil replied: I don't have experience with AMD GPUs.

Wenhao Sun addressed Veronika Juraskova, Takuya Taniguchi and Venkat Kapil: I am wondering if you can give some perspective on what kinds of problems you think are most addressable by machine-learning potentials, for example in the solvation space. In my mind, there are several tiers of questions you can tackle with MLIP-MD simulations – mechanisms, dynamical properties (diffusivity, forces), structure, and energy. What do you think are addressable? From my perspective, I think it depends on how quantitative of a result you need. Mechanism is often quite qualitative, and even with inaccurate dynamics or energies, you might still be able to broadly make assertions on mechanism (*e.g.*, rate-limiting step). I feel like dynamics are a bit more quantitative, but still you can get emergent dynamical phenomena with slightly inaccurate potentials. I feel like structure (bond lengths, bond angles), should require highly accurate potentials.

And finally, I think energy is the most difficult and unreliable to get accurately. Would you agree? I think a counter-argument might be that MLIPs are trained on structure and energies, so maybe they would be the most reliable. But I'm curious on your perspective.

Veronika Juraskova answered: The conclusion on which question is the most addressable by MLP-MD is not simple to answer and will be system-dependent. Machine learning-based potentials we train allow us to reproduce the energy and forces of the ground truth method for a given structure. Consequently, the MLP can be only as accurate as the reference. In solution, the chemical observables are influenced by the interplay between non-covalent interactions with solvent molecules, entropy and final temperature effects. The properties which are most sensitive to the interactions with chemical environments will be therefore most influenced by the accuracy of the potential and the quality of their description. These types of properties will especially benefit from the MLP-MD, as it allows to reach high *ab initio* accuracy together with the long sampling.

Takuya Taniguchi responded: Thank you for your interesting question. I agree with your perspective. Organizing the applications hierarchically is indeed helpful for understanding. For example, in the solvation space, I believe the advantages of MLIPs can be utilized in the following types of simulations:

(i) Simulations of large-scale solvent–solute systems.

(ii) Tracking long-time dynamical processes (*e.g.*, solvation and desolvation processes of ions).

(iii) Predicting structural changes of solutes under various solvent conditions.

These applications leverage the computational efficiency of MLIPs, allowing for the study of complex systems and processes that might be computationally prohibitive with traditional *ab initio* methods. The ability of MLIPs to handle large systems and longtime scales makes them particularly suitable for these types of investigations in the solvation space.

Venkat Kapil replied: In theory, state-of-the-art machine learning interatomic potentials can be applied to study both structural and dynamic properties, provided the underlying potential energy surface is accurate. Therefore, if you can identify an accurate electronic structure theory level (*e.g.*, density functional theory or a higher-level method) and produce a representative dataset, you should be able to address the problems you mentioned. The main challenges include ensuring confidence in the electronic structure theory level and having sufficient computational resources to train on an adequate set of configurations, including transition states for chemical reactions. I hope that with the fine-tuning approach presented here, we can minimize the number of required training configurations and make accurate electronic structure theory calculations more affordable.

Matthew R. Ryder communicated: Many papers in this session highlight computational efficiency gains through machine learning. I wonder where the trade-off lies between computational speed and the level of accuracy required for real-world applications, especially in systems where disorder and thermodynamic effects are significant.

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.

