Shifting computational boundaries for complex organic materials

Methodology adapted from data science sparked the field of materials informatics, and materials databases are at the heart of it. Applying artificial intelligence to these databases will allow the prediction of the properties of complex organic crystals.

R. Matthias Geilhufe, Bart Olsthoorn and Alexander V. Balatsky

Materials used today often have a complicated structure that makes their properties hard to predict, and so those with good specific functions are usually discovered serendipitously. And, while our computational capabilities are growing rapidly, there are still many situations that are too complex for accurate simulation. As a result, we have a gap between the technological demand for complex materials with desired properties and materials and properties that can be reliably calculated using modern ab initio methods. We focus on this tension in the context of organic materials, for two reasons. First, carbon-based compounds show an infinite configuration space and, as a result, an enormous potential for functional materials. Second, owing to the complexity of the compounds, the computational gap is particularly large for this class of materials. Organics have become a crucial part of modern technology, medicine and chemistry, with wide-ranging applications, such as organic light-emitting diodes, solar cells or explosives, and have been discussed in the context of molecular qubits and spin–liquid physics.

The field of materials informatics deals with the application of methodology adapted from computer science to materials research. With the exponential increase in computational power and storage capacity, numerous materials databases have emerged in recent decades, often encouraged by focused national or international strategic funding initiatives. Such databases are an exciting domain for data mining and machine learning.

Within this trend, we focus on recent computational databases for carbon-based materials (Table 1); for a broad overview of existing materials databases, see, for example, refs. 4–6. A prominent series of datasets storing calculated molecular properties is available at http://quantum-machine.org/datasets/. Calculated properties of crystalline organic materials are available via the Organic Materials Database (https://omdb.mathub.io). Properties of organic polymers can be found at Polymer Genome (https://www.polymergenome.org/). Subsequent work has shown that datasets from these resources are valuable resources to train machine learning algorithms.

Data mining describes the quick exploration for materials with desired target properties within a materials database. This can be achieved by constraining the search space by specific parameters or by using pattern matching. One can use established heuristics where a material’s property is related to specific patterns in the available data, such as the band structure. For example, topological semimetallicity corresponds to the occurrence of nodes in the band structure, or superconductivity is linked to peaks in the density of states at the Fermi level. Dedicated algorithms that search for similarities to specified initial patterns implemented into materials databases5–10 provide a versatile tool to identify materials in almost no time, as has been shown for organic Dirac materials and topological materials11,12, superconductors10, energy materials13, magnetic materials14 and quantum sensors15.

Given the great success of data mining on materials databases for organics, a natural question arises: why can’t we continue along this promising path forever? One of the big bottlenecks constraining this route is given by the intrinsic complexity of organic materials. The space of computationally feasible materials strongly depends on the level of approximation used in the simulation (Fig. 1a). For example, density functional theory (DFT) is currently the most widely used first-principles approach to calculate materials properties. Although reports towards linearly scaling DFT approaches exist16, the computational costs of conventional DFT codes increase as $O(N^3)$, where $N$ is the number of atoms. Regarding all materials containing carbon and hydrogen in the Crystallography Open Database17, their unit cell sizes can be approximated to follow a log-normal distribution with a mean of 222 atoms in the unit cell (Fig. 1b).

While an organic structure is typically very complicated, it also shows recurring local patterns with similar behaviour of

| Database                  | URL                        | Specifications                                                                 | Reference |
|---------------------------|----------------------------|-------------------------------------------------------------------------------|-----------|
| Organic Materials Database| https://omdb.mathub.io     | Organic molecular crystals, metal–organic frameworks; band structure, density | 18        |
|                           |                            | of states, magnetism, pattern matching, machine learning tools                |           |
| Polymer Genome            | https://www.polymergenome.org/ | Experimental and computational data of polymers; machine learning tools         | 31        |
| QM7, QM7b, QM8, QM9       | http://quantum-machine.org/datasets/ | Small organic molecules; geometries minimal in energy, corresponding harmonic frequencies, dipole moments, polarization, energies, enthalpies | 32        |
local functional properties such as charge density, spin density and orbital density, or strain profiles. Machine learning models trained on computationally feasible crystal structures could provide reliable predictions of such properties in materials outside the reach of ab initio methods. Such recurring patterns also play a central role in machine learning descriptors — numerical representations of the structures — for polymers, where the material is built of large molecules composed of many repeating subunits. The local environment around an atom in real space is most commonly used to predict material properties, while non-local representations are used to include long-range physics.

The structure–property relationship of organic molecules was first captured successfully by using handcrafted descriptors such as the Coulomb matrix in 2011. This study initiated many works introducing new descriptors that have desired properties, such as rotational symmetry or uniqueness of the descriptor. Recently, there have been two updates to this scheme. First, using more sophisticated techniques from data science, the construction of explicit descriptors is usually no longer a separate step, but rather integrated into the regression model itself. Two examples are SchNet, a continuous-filter convolutional neural network, and SOAP, a similarity kernel based on the integration of atomic positions. Second, the next step beyond scalar properties is the prediction of higher-dimensional objects, such as the density of states, band structure, or charge and magnetization density. This direction is currently actively worked on and progress is promising, for example, with respect to reliable charge density predictions. Supervised machine learning algorithms approximate an unknown or computationally expensive map between input and output vectors. As such, they provide predictive power within the range of the given data. However, attempting extrapolation usually fails. For certain machine learning methods, for example, Bayesian inference, a large uncertainty in the prediction can provide a sign of extrapolation. Informally speaking, the failure of extrapolation in neural networks can be understood by the observation that deep nets are essentially performing polynomial regression with high-degree polynomials.

Fig. 1 | Organic molecular crystals and metal–organic frameworks typically have highly complex unit cells with hundreds of atoms present. Conventional ab initio approaches, in particular taking into account strong correlation corrections, are not feasible after a certain complexity is reached. Hence a new methodology is required. a, Outline of the range of current methodology and the number of atoms (N\text{atoms}) per unit cell for certain experimental crystal structures. DFT-based codes are feasible for structures with several thousand atoms. Computationally feasible structures are smaller by at least one order of magnitude for standard implementations incorporating dynamical mean field theory (DMFT) or self-energy corrections using the GW approximation. b, The log-normal distribution of organic materials in the Crystallography Open Database. The mean number of atoms is e^\mu = 223 atoms with a standard deviation \sigma of the natural logarithm of about 0.67.

Fig. 2 | Machine learning techniques are designed for data regression and typically fail when it comes to extrapolation. However, certain machine learning frameworks trained on local environments in materials with feasible unit cell sizes can be applied to predict the properties of complex materials with similar local properties. Using the example of the bandgap, we show the N\text{atoms}-resolved mean absolute error for an ensemble of machine learning models inside and outside the range contained in the training set. The orange shaded region illustrates the standard deviation. As the model is based on local atomic environments, an ‘extrapolation’ towards more complex materials remains fairly stable with a well-behaved standard deviation up to 500 atoms in the unit cell. COD, Crystallography Open Database.
polynomials. The flexibility of capturing any oscillations within the dataset provides predictive power, but diverges outside the known data points. However, by training a model on reoccurring local environments and their properties in small organic crystals, an application to complex materials does not necessarily fall into this class. In fact, a sufficiently large dataset containing simple organic molecular crystals will provide the majority of local atomic neighbourhoods present in a dataset of complex organic crystal structures.

We show this reasoning on the example of the bandgap in non-magnetic organic crystals, where we follow the approach of ref. 1 We train 100 SchNet models (for hyperparameters refer to ref. 1) on identical training, validation and test data. However, each training of the model starts by a random initiation and therefore ends up with slightly different learnable parameters. We can benefit from this behaviour by discussing a collection of trained models, and taking the mean and standard deviation of the bandgap prediction. The training and validation dataset contains only materials with fewer than 112 sites, whereas the test set only contains larger materials. The SchNet infrastructure predicts a site contribution based on each local site environment in the unit cell before a final pooling layer gives the scalar result. For intensive properties, such as bandgap in this work, an average is used. Our model trained on 25,313 materials taken from the Organic Materials Database achieves a mean absolute error of approximately 0.4 eV, which is a reasonable accuracy, as organic crystals are typically large bandgap insulators with an average gap of around 3 eV (ref. 13). In Fig. 2, we show the evolution of the mean absolute error and the standard deviation of the mean absolute error achieved by our ensemble of machine learning models resolved for the number of atoms in the unit cell. While the model is trained on materials with fewer than 110 atoms in the unit cell, we can see that the general trend of our predictive power is also maintained in region of up to 150 atoms in the unit cell. We also evaluate all 100 models on 213,681 materials in the Crystallography Open Database, and show that the standard deviation in Fig. 2 remains fairly constant up to about 500 atoms in the unit cell. If we were in a pure extrapolation regime, the predictions of neural networks typically grow wild and as such the standard deviation would increase rapidly. This result shows a promising direction towards predicting materials properties outside the range of materials that can be described with conventional ab initio methods.

Unsupervised machine learning allows identification of the inner structure in data, for example, by dimensionality reduction and clustering. It also provides the prospect of generative modelling, providing algorithms that generate authentic fake data following a given statistical distribution. Prominent algorithms are variational autoencoders17 or generative adversarial networks18. Recently, such models have been applied more frequently to complex materials and drug design19,20. This development is a huge improvement towards the hard task of computationally determining an unknown stable molecular or crystal structure. Closing the loop by combining the prediction of stable structures with the prediction of desired properties represents a huge advance to materials design, substantially extending the available search space.

In summary, we see an increasing interest in highly complex but fascinating organic crystals. This interest is fostered by an increasing computational power along with the adoption of new tools that have made it possible to conduct sophisticated materials simulations at the same time as synthesis and experiment. The success of the current approaches for using materials informatics largely rests on the established paradigm of scaled-up ab initio methods: structure −→ properties −→ database −→ end user and applied machine learning tools. Perhaps, with the new tools we have described that use ensembles of various machine learning models for various materials properties we will see the emergence of a new phase of computational complex materials research. As such, the outcome of machine learning algorithms could foster the generation of a second generation of databases involving a much larger scale of feasible materials. Such a workflow would be of a different form: (artificial intelligence −→) structure −→ artificial intelligence −→ properties −→ database −→ end user. This ‘artificial knowledge’ has the potential to predict materials properties beyond the known constraints of conventional ab initio tools. It would, for example, build the bridge between materials informatics and organized functional biomolecules21. While such a methodology needs to be developed and evaluated in more detail, we believe that this route is a promising candidate towards exascale materials simulations.

R. Matthias Geilhufe✉1,2✉, Bart Olthoorn and Alexander V. Balatsky

Departments of Chemistry and Materials Science, University of Connecticut, Storrs, CT, USA.

References

1. Gaita-Ariño, A., Luis, F., Hill, S. & Coronado, E. Nat. Chem. 11, 301–309 (2019).
2. Shimizu, Y., Miyagawa, K., Kanoda, K., Maestas, M. & Saito, G. Phys. Rev. Lett. 91, 107001 (2003).
3. Lefayd, R. A. Organometallics 33, 1084–1099 (2014).
4. Schleder, G. R. et al. J. Phys. Mater. 2, 032001 (2019).
5. Hinamen, L., Geurts, A., Foster, A. S. & Rinke, P. Adv. Sci. 6, 1900808 (2019).
6. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Nat. Commun. 9, 8887 (2018).
7. Geilhufe, R. M. & Olsthoorn, B. Phys. Rev. B 102, 205134 (2020).
8. Haan, T. D. et al. J. Appl. Phys. 128, 175104 (2020).
9. Borisyov, S. S., Olthoorn, B., Gedik, M. R., Geilhufe, R. M. & Balatsky, A. V. npj Comput. Mater. 4, 46 (2018).
10. Geilhufe, R. M., Borisyov, S. S., Kalpachchi, D. & Balatsky, A. V. Phys. Rev. Mater. 2, 024002 (2018).
11. Geilhufe, R. M., Bouhon, A., Borisyov, S. S. & Balatsky, A. V. Phys. Rev. B 95, 041103 (2017).
12. Geilhufe, R. M., Borisyov, S. S., Bouhon, A. & Balatsky, A. V. Sci. Rep. 7, 7228 (2017).
13. Zhang, L., Chen, Z., Su, J. & Li, J. Renew. Sustain. Energy Rev. 107, 554–567 (2019).
14. Hellvik, J., Pérez, R. D., Geilhufe, R. M., Männson, M. & Balatsky, A. V. Phys. Rev. Mater. 4, 024009 (2020).
15. Geilhufe, R. M. et al. Phys. Status Solidi Rapid Res. Lett. 12, 1800293 (2018).
16. Fonseca Guerra, C. F., Steijgers, J. G., te Velde, G. & Baerends, E. J. Chem. Acc. 99, 391–403 (1998).
17. Grañulis, S. et al. Nucleic Acids Res. 40, D420–D427 (2012).
18. Borisyov, S. S., Geilhufe, R. M. & Balatsky, A. V. Phys. Rev. B 112, 017101 (2017).
19. Haan, T. D., Mannodi-Kannam, A. & Rumaspasad, R. Phys. Rev. B 92, 041106 (2015).
20. Grisafi, A. & Cerotti, M. J. Phys. Chem. 151, 204105 (2019).
21. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Phys. Rev. Lett. 108, 038001 (2012).
22. Schütt, K. et al. Adv. Neural Inf. Process. Syst. 30, 991–1001 (2017).
23. Bartók, A. P., Kondor, R. & Csányi, G. Phys. Rev. B 87, 184115 (2013).
24. Tiubaldi, M. & Múnoz-Gachó, T. Phys. Rev. Lett. 125, 206401 (2020).
25. Gong, S. et al. Phys. Rev. B 100, 184103 (2019).
26. Kinross, D. P. & Welling, M. Preprint at https://arxiv.org/abs/1411.1440 (2013).
27. Goodfellow, I. et al. Adv. Neural Inf. Process. Syst. 27, 2672–2680 (2014).
28. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Science 361, 360–365 (2018).
29. Jackson, N. E., Webb, M. A. & de Pablo, J. J. Curr. Opin. Chem. Eng. 23, 106–114 (2019).
30. Simon, A. J. et al. Nat. Chem. 12, 204–212 (2019).
31. Haan, T. D. et al. Sci. Data 7, 160012 (2016).
32. Ramakrishnan, R., Dvál, P. O., Rupp, M. & von Lilienfeld, O. A. Sci. Data 6, 140022 (2014).

Acknowledgements

We are grateful for collaboration and discussions with J. Hellvik and S. S. Borisyov. We acknowledge funding from the ERC synergy grant ‘HERO’ (number 810451), the University of Connecticut, VILLUM FONDEN via the Centre of Excellence for Dirac Materials (grant number 11741), the Knut and Alice Wallenberg Foundation (2019.0068) and the Vetenskapsrådet (number 2017.03997). We acknowledge computational resources from the Swedish National Infrastructure for Computing (SNIC) at the Centre for High Performance Computing (PDC), the High Performance Computing Centre North (HPC2N) and the Uppsala Multidisciplinary Centre for Advanced Computational Science (UPPMAX).

Competing interests

The authors disclaim no competing interests.

Published online: 12 January 2021
https://doi.org/10.1038/s41567-020-01135-6