How the shape of chemical data can enable data-driven materials discovery

Jacqueline M. Cole,[1,2,3,4*]

[1] Cavendish Laboratory, Department of Physics, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, UK.

[2] ISIS Neutron and Muon Facility, STFC Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot, OX11 0QX, UK.

[3] Department of Chemical Engineering and Biotechnology, University of Cambridge, West Cambridge Site, Philippa Fawcett Drive, Cambridge, CB3 0AS, UK.

[4] Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK

* Correspondence: jmc61@cam.ac.uk (J.M. Cole)

Abstract

Chemical data have been created from many different origins. The chemicals themselves tend to be synthesized out of curiosity or an industry-led need. Their materials characterization and development for functional applications generate cognate data about their structures and properties. Chemical structures and properties may also be computed ahead of their physical creation. The collation of all this chemical information affords a 'chemical space' that encapsulates a rich and diverse set of data. This paper considers the shape and size of this chemical space, and that of its various sub-domains; and how the relative availability of its structure and property information governs what type of questions one should ask of the data, and what type of machine learning should be applied, to discover a new material. Application examples of machine-learning methods that produce predictive models for data-driven materials discovery are discussed.

Keywords: chemical space, machine learning, structure-property relationship, data-mining

Data-driven materials discovery

The Materials Genome Initiative [1] has played a key role in promoting data-driven materials discovery. Its industrial motivations are clear, presenting a need to overcome the 20-year average timeline of the 'molecule-to-market' pipeline. Industry tends to progress a lot faster in other areas of its technological development and so materials discovery is a key bottleneck to innovation and thus the economy. Materials discovery is still largely realized by serendipity; slightly better are 'trial and error' approaches, while a systematic approach is really needed. Fortunately, the rise of artificial intelligence, high-performance computing, and open-data government regulations has provided a 'big data' opportunity for systematizing materials discovery. The aim is to realize a full 'design-to-device' pipeline of data-driven materials discovery for any given application [2-6].

The success of data-driven materials discovery relies on having enough of the right type of data. Historically, crystal structures have provided the largest sources of materials data, given that they have been collated to form databases such as the Cambridge Structural Database (CSD) [7], the Inorganic Crystal Structure Database (ICSD) [8], the protein databank (PDB) [9], and the Crystallography Open Database (COD) [10,11]. These databases enabled early examples of data-driven materials discovery, by concerting its structural information with quantum-chemical calculations. Examples include the use of the CSD to discover new classes of organic non-linear optical materials [12] and new light-harvesters for dye-sensitized solar cells [13]. During the same era, large computational databases began to be designed and constructed that predicted new materials for bespoke applications. For example, a computationally generated database of conjugated polymers was created for prospective organic photovoltaic applications [14]. Another database of this ilk created a vast array of prospective light-emitting materials; that work went one step further by experimentally validating its lead candidates to realize materials discovery for light-emitting diode (LED) applications [15]. High-throughput computation was also used to amass large databanks of structure-property information for more general use, a notable example being the Materials Project [16]. These examples are in fact representations of the 'big data' approach before 'big data' emerged as a recognized field of research.

The implementation of government-regulated mandates to enforce the open access of data has been instrumental in enabling the 'big data' opportunity for scientific pursuit. In materials science, structure and property information about chemicals are strewn across the scientific literature in highly fragmented forms. The prospect of collating such data has become a reality given the shift to open access of data. This has motivated efforts to mine chemical data from the literature. This follows early efforts to mine chemical text which were focused on small-scale data harvesting, from particular sections of text in a paper, using natural language processing (NLP) methods; notable examples are ChemicalTagger [17] and LeadMine [18]. These developments were needed since the textual language of the scientific literature is far too specialized to be mined using generic text-mining tools, such as CoreNLP [19] and SpaCy [20], that have been developed in computer science. A clear need to produce a high-throughput 'chemistry aware' text-mining tool to auto-generate large chemical databases led to the development of the open-source ChemDataExtractor tool [21]. This tool has already been used to create large databases that have enabled the discovery of new light-harvesting materials for dye-sensitized solar cells [22] and magnetic refrigerants for hydrogen storage [23,24]. ChemDataExtractor has also been used to auto-generate new banks of materials characterization data; a notable example being a UV/vis absorption spectral databank [25]. Image-based materials-characterization data can also be extracted using ChemDataExtractor in conjunction with a high-throughput image-mining tool, ImageDataExtractor, which has been developed to quantify electron microscopy data [26]. A similar tool has been developed to analyze the shape and size of nanoparticles [27].

In principle, one could mine any type of chemical data, although it is important to remember that the

literature being mined will normally contain processed data. Such data are generally the most useful to a data-science practitioner since humans have already transformed raw data into a readily interpretable form. Processed chemical data will take many forms owing to the processes by which they are created to study a material. Figure 1 illustrates the data conveyor belt for the creation of experimental data and the use of data science to afford a 'molecule-to-market' pipeline.

Databases themselves do not lead to materials discovery. Rather, patterns in such databases that pertain to structure-property relationships tend to be the source of data-driven materials discovery. Historically, such patterns were harnessed by manual surveys of chemical databases. Yet, the widening availability of machine-learning and statistical inference methods for the materials-science community has revolutionized this area over the last few years. This is because such methods can automatically find patterns in data concerning structure-property relationships. This affords them the required predictive quality for data-driven materials discovery.

Nonetheless, each machine-learning (ML) method has particular data needs. On the one hand, an ML method may need to be trained using a collection of reference data that has been manually labeled with the features that are ultimately sought automatically from a much wider set of (unlabeled) data. ML methods that need training data are called supervised ML methods. Those that need no training are called unsupervised ML methods. Semi-supervised ML methods need training data but not as much as supervised ML methods since they can then learn progressively from the data to which they are ultimately applied; this dynamic learning process is known as bootstrapping owing to its statistical basis of random sampling by replacement. On the other hand, each type of ML method needs a certain amount of data to function satisfactorily, according to factors such as the number of parameters required to determine a model by a given ML method and the level of noise in the data. This relates to the intrinsic need for a minimum data:parameter ratio. Too few parameters would result in an ML model underfitting, thereby rendering it with poor predictive performance. So, ML-methods based on linear regression can meet its data:parameter ratio needs fairly easily, while a parameter-hungry ML method such as a neural network, would need far more data than most other ML methods. Meanwhile, a higher data:parameter ratio would be needed for an ML method to model particularly noisy data. So, a lot more data may be required in such circumstances; alternatively, more extensive parameterization may prove sufficient to uncover the more latent patterns within the data.

The shape and size of chemical data that are available to a given ML task also vary massively. Each field of chemistry has its own characteristic subset of chemical space which may be spare or rich, similar or diverse, homogeneous or heterogeneous. Chemical data may also comprise many types of information: from synthesis, metrology, property or device testing. An ML model may need to be fashioned accordingly. For example, if the chemical data for a given problem are largely structural, then a different approach is taken compared to a problem whose chemical data contain both structure and property information. The size of the chemical dataset involved will also have a bearing on the strategy for data-driven materials discovery. To this end, the next section of this paper surveys the shape and size of various forms of chemical data. The ML choices for data-driven materials discovery are then exemplified, within a framework where chemical data are conditioned against the level of structure and property information that lies within them.

The shape and size of chemical data

Chemical space is highly diverse, and its density varies widely across different classes of chemicals. This heterogeneity exists because the creation of new chemicals is governed by external influences that are largely unsystematic and full of human bias. This is best explained by example. Figure 2 considers how the growth of chemical space over the last 20 years has been driven by technological need, life concerns, and the availability of chemical structures and key synthetic reactions.

The rise and fall of an emerging technology or world need will propel or abate the growth of certain fields of chemistry. For example, Figure 2a shows how the range of available nanomaterials grew rapidly in 2000 when the field of nanotechnology was becoming established; as the field has matured, it remains a large field in 2020, but its growth has waned since 2000. Battery research has been strong for several decades, but it has grown substantially over the last 20 years owing to increasing need to create new materials for battery-powered technologies that will offset climate change; the rise of electric cars is a particularly visible example of this need. Another growth area for chemistry in the energy sector is photovoltaic research. Overall, this field has proliferated over the last 20 years, but the demographics of the types of solar cells being studied have changed massively over this time. Organic photovoltaics (OPVs) and dye-sensitized solar cells (DSCs) were very popular in 2000, but the discovery of perovskite-based solar-cells (PSCs) has transformed this field of chemistry. As such, PSCs are now one of the largest growth areas of chemical research, while studies on OPVs and DSCs have waned in the fashion shown in Figure 2a. Other current areas of chemical research that are now very popular barely existed in 2000; Figure 2a provides examples that include topological insulators, metal- and covalent-organic frameworks (MOFs and COFs), quantum materials, and nanomaterials.

The research and development associated with these functional-material applications has created a wealth of chemical and property data, in the form of materials fabrication, characterization, and metrology information. These data have found their way to the literature and NLP-based tools have automatically collated them to form materials databases; a recent example being a battery materials database [28] that was auto-generated using ChemDataExtractor [21]; it contains over ¼ million records of chemical names and their associated battery properties: voltage, conductivity, capacity, Coulombic efficiency, and energy. ¼ million data records is an important milestone for ML needs since this is considered to be a sufficient amount of data for all but the most complex and deep neural networks.

Meanwhile, large datasets have been manually curated over the last 20 years, in the absence of 'chemistry aware' NLP tools, to address life needs; *cf.* Figure 2b. The mapping of the human genome [29], which is now in its complete (X-chromosome) form [30], contains over 3 million base pairs. This compares with databases of natural products that have grown over a similar timeframe with 'Super Natural' database of 50k compounds being released in 2006 [31], rising to 326k compounds in 2015 with 'Super Natural II' [32]. An amalgamation of multiple natural products databases led to the COCUNUT databank of 400k compounds in 2020 [33].

The relative scales of chemical data are important. In comparison to crystallography databases, databases about life are modest in size; *cf.* Figure 2c. For example, the Cambridge Structural Database alone is far larger than the entire known chemical set of natural products. This stands to reason since chemical structures can feature any element of the periodic table, while the fundamental building blocks of life are confined to a characteristic set of amino acids. The far greater diversity of materials chemistry points to a seemingly infinite materials genome [34]. The need for crystal structures has also influenced the growth of chemical data. The importance of landmark crystal structures for developing a new field cannot be understated given that so many Nobel prizes have been awarded to crystallographers [35]. This is partly because crystal structures are considered to be the 'gold standard' of materials characterization, generally being able to confirm unambiguously the chemical structure of a material. So, crystal structures are highly sought. Indeed, the CSD realized 1 million crystal structures last year. Yet, such data can only be obtained if a chemical crystallizes. Accordingly, crystallography databases feature distinct gaps in their survey of chemical space where certain types of chemicals will not crystallize.

This need for crystallization could be circumvented by computing crystal structures instead. Such calculations are usually performed via density functional theory (DFT). Purely computational databases have evolved recently, given the massive growth in opportunities for high-performance

computing. Such databases contain crystal and molecular structures of chemical compounds with impressive tallies: Open Quantum Materials Database (OQMD) (816k) [36,37], AFLOWLIB (3.3M) [38], the Materials Project (710k) [16], and NOMAD which features 50M total energy calculations [39]. That said, there is a need for standardization in computationally generated structural data. In terms of computed crystal structures, the figures-of-merit that have long been used to validate experimentally determined crystal structures can hopefully help evolve this standardization for computed structures. Tools that enable interoperable materials databases will also help standardization, a prominent example being https://www.optimade.org/.

The Inorganic Crystal Structure Database started to accept computed structures into their hitherto experimental databank in 2017, in cases where they are deemed to be sufficiently accurate [40]. This amalgamation of experimental and computed crystal structures represents an important milestone because it facilitates joint experimental and computational approaches to data science. Since chemical and physical properties can be predicted from computed crystal structures, this amalgamation also opens the purely structural nature of the ICSD to a structure-property chemical space. This structure-property framework is crucial to the success of ML-based data-driven materials discovery, as will be illustrated in the next section.

The computational determination of crystal structures also overcomes the synthetic limitations of chemical space. Chemists are practically restricted by the synthetic routes that are available to them. This imposes a severe limitation on the diversity of chemical space that can be experimentally realized. For example, the top five most common synthetic reactions for medicinal chemistry, shown in Figure 2d, represent about half of all synthetic reactions carried out in this field [41]. Computation carries no such restriction, and so it can be used to artificially create any compound, computer-resource permitting. This lack of restriction for computation also means that it can lead the way in the systematic mapping of chemical space. This might be via an exhaustive grid-style mapping of chemical space, as a function of a certain parameter, such as the number of atoms; at least for molecular structures, this has been achieved for small molecules, e.g., QM9 [42]. There are potential issues of validation associated with such an approach, although reality checks could be imposed on the computed structures by comparing, or even anchoring, them to cognate experimental data where they are available at certain points in the grid. Even with such validation, though, it is not yet practical to consider computationally mapping the molecular or crystal structures of the entirety of chemical space, given that it is reputed that $> 10^{80}$ chemical compounds could exist in the Universe [2].

Engaging ML with chemical structure and property data to afford materials discovery

The success of data-driven materials discovery is contingent on asking the right questions of the data. In turn, the choice of questions depends upon the type of chemical space that is available. Such data typically involve chemical structures or their properties or both. This is because structure-property relationships lie at the origin of most patterns in chemical space that ML can use to predict new chemical materials or properties. It is thus helpful to condition our enquiries into four overarching questions that are based on various permutations of structure and property. Within this scope, we now exemplify how ML can engage with chemical data to afford predictive models for materials discovery. The type of chemical data, ML method for each example application is also summarized in Figure 3.

Given the structure of known materials, what are the structures of other materials?

Crystallography databases are particularly helpful as a data source for ML to answer this type of query. For example, an ML model that predicts the dimensionality of hydrogen-bonded networks in crystal structures has been created by using 64,084 hydrogen-bond networks that capture this dimensionality in known crystal structures from the CSD together with a support vector machine (SVM) classification method [43]. Such a model could be useful to design the supramolecular chemistry of

materials, while the authors of this ML study highlighted the prediction of mechanical behavior in materials as a specific application [44]. 8,050 crystal structures of co-crystals in the CSD have also been used, together with the same number of deliberately invalid co-crystals, to help train an ML model, in the form of an artificial neural network, that predicts the propensity of two chemicals to co-crystallize [45]. The study afforded a reliable predictor for co-crystallization, which is especially important for drug manufacture.

Given the structure and properties of known materials, what are the properties of other known material structures?

Databases of crystal structures can be augmented by their cognate property information using quantum-chemical calculations. Structure-property relationships unfold as patterns in the data wheresoever they exist. This affords ML an ideal opportunity to use these correlated data to train a model that predicts a given property in other structures. For example, 260,092 band gaps for organic crystal structures have been predicted [46] using two ML models, based on ridge-regression or deep-learning, which were applied to 260,092 experimentally-determined crystal structures from the Crystallographic Open Database (COD). These ML models were formulated using 12,500 crystal structures and their band gaps, both of which were computationally generated using electronic-structure calculations. In another study, a deep-learning model was used to predict the hardness and fracture toughness of materials [47] for 120,000 structures in the Materials Project. This model was trained and tested using 8,033 computationally generated crystal structures and elastic property data from the Materials Project [16] together with an empirical model for hardness and toughness [48].

Given the structure and properties of known materials, what are the structures and properties of new materials?

The prediction of *new* materials and their properties requires a *generative* ML approach. This is the mainstay of Generative Adversarial Networks (GANs) [49] and Variational Auto-Encoders (VAE) [50]. VAEs have shown success in generating thousands of new crystal structures that are suited to a given property within a matter of hours. A crystal graph convolutional neural network (CGCNN) [51] is then used to predict the properties of these new structures. In one recent example of such a VAE-CGCNN pipeline [52], the VAE receives a known material as input and generates thousands of new crystal structures that are similar to the input, e.g., an atom has been substituted or bond geometry has changed from the input. This 'morphing' behavior is achieved by ML using a 'latent space' that is trained from known crystal structures that exhibit the property of interest; in the subject study, the training data comprised 7,189 structures from the Materials Project. The CGCNN then calculates certain properties for each new structure, which in this case were: formation energy, total energy, band gap, bulk modulus, shear modulus, Poisson ratio, refractive index, and dielectric constant. The CGCNN for the subject study was also trained using data from the Materials Project.

Given the properties of known materials, what are the properties of other materials?

One may not always have the luxury of crystal structure data. However, the chemical formula of a material still offers insights into its structure, and how that relates to its properties. Accordingly, ML models can fashion structure-property relationships using the chemical constituents and chemical compositions of a set of materials together with their property data. A recent example of this approach is the data-driven discovery of a new magnetic refrigerant, HoB_2 [24], whose magnetocaloric effect (MCE) operates at the liquefaction temperature of hydrogen (20.3 K). Thus, it could store hydrogen for energy fuel applications. Gradient boosting with Bayesian optimization ML methods were used to find a material that exhibits MCE around this temperature, using a database comprising 39,822 chemical formulae and their Curie and Néel phase-transition temperatures that were mined from the literature using ChemDataExtractor [23]. The chemical constituents and compositions of the formulae

were used to create feature vectors that were combined with reported values of the changes in entropy due to MCE, ΔS_M, and applied field, ΔH which are related by a Maxwell equation. These were used to train an ML model which employed a gradient-boosted tree algorithm and Bayesian optimization. The ML model was then used to predict ΔS_M values of 818 chemicals in the magnetic materials database that hosts their compositions and Curie temperatures [23]. The non-toxic material alloy whose ΔS_M value is closest to the hydrogen liquefaction temperature, and contains heavy rare-earth elements, Gd-Er, was selected. This was HoB_2, whose MCE characteristics were then experimentally validated.

Concluding remarks

The world is reacting to a timely opportunity for data-driven materials discovery, to realize the objectives of the Materials Genome Initiative[1,53]. Machine learning lies at the fore, but it cannot succeed without the right type and amount of chemical data. Materials databases containing chemical and property data can now be auto-generated by mining and collating all of the highly fragmented data that exists about a target chemical application from the literature. This database auto-generation process uses artificial intelligence, in the form of natural language processing, which has been made 'chemistry-aware' given that the success of text-mining is so domain specific. Such databases complement those which have been constructed manually over the years, especially crystallography databases which are particularly helpful for materials discovery. These structural databases need to be considered within the wider scope of chemical data. Thereby, the shape and size of chemical data have herein been surveyed. Its highly diverse and heterogeneous form, and variable density across different classes of chemicals, has been explained in terms of the unsystematic influences by which the growth of chemical data is driven. The use of electronic structure calculations to help regularize and extend the unsystematic nature of chemical data has been described. The manner by which ML can engage with chemical data has thence been exemplified, bearing in mind the framework in which questions of the data need to be conditioned to solve a given materials problem, given the relative availability of structure or property data. Ultimately, structure-property relationships lie at the foundations of patterns in chemical data and ML models are ideally suited to predict new materials chemistry based on these patterns.

Acknowledgements

J.M.C. is grateful for the BASF/Royal Academy of Engineering Research Chair in Data-Driven Molecular Engineering of Functional Materials, which is partly supported by the STFC via the ISIS Neutron and Muon Facility.

References

[1] Materials Genome Initiative for Global Competitiveness; National Science and Technology Council, Office of Science and Technology Policy: Washington, DC, 2011

[2] Cole, J. M. (2020) A design-to-device pipeline for data-driven materials discovery. Acc. Chem. Res. 53, 599–610

[3] Himanen, L. et al. (2019) Data-Driven Materials Science: Status, Challenges, and Perspectives. Adv. Sci. 6, 1900808

[4] Agrawal, A. and Choudhary, A. (2016) Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. APL Mater. 4, 053208

[5] de Pablo, J. J. et al. (2019) New Frontiers for the Materials Genome Initiative. npj

Computational Materials 5, 41

[6] Alberi, K. et al. (2019) The 2019 Materials by Design Roadmap. J. Phys. D: Appl. Phys. 52, 013001

[7] Groom, C. R. et al. (2016) The Cambridge Structural Database. Acta Cryst. B72, 171-179

[8] Bergerhoff, G. et al. (1983) The Inorganic Crystal Structure Data Base. J. Chem. Inf. Comput. Sci. 23, 66-69.

[9] Berman, H. M. et al. (2000) The Protein Data Bank. Nucleic Acids Res. 28, 235-242

[10] Grazulis, S. et al. (2009) Crystallography Open Database – an open-access collection of crystal structures. J. Appl. Cryst. 42, 726-729

[11] Gražulis, S., et al. (2012) Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Res. 40, D420-D427

[12] Cole, J. M. and Weng, Z. F. (2010) Discovery of High-performance Organic Non-linear Optical Molecules by Systematic 'Smart Material' Design Strategies. Adv. Mater. Res. 123–125, 959–962

[13] Cole, J. M. et al. (2014) Data Mining with Molecular Design Rules Identifies New Class of Dyes for Dye-sensitised Solar Cells. Phys. Chem. Chem. Phys. 16, 26684– 26690

[14] Hachmann, J. et al. (2011) The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. J. Phys. Chem. Lett. 2, 2241–2251

[15] Gómez-Bombarelli, R. et al. (2016) Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nat. Mater. 15, 1120-1127

[16] Jain, A. et al. (2013) The Materials Project: A materials genome approach to accelerating materials innovation. APL Mater. 1, 011002

[17] Hawizy, L. et al. (2011) ChemicalTagger: A Tool for Semantic Text-mining in Chemistry. J. Cheminf. 3, 17

[18] Lowe, D. M. and Sayle, R. A. (2015) LeadMine: A Grammar and Dictionary Driven Approach to Entity Recognition. J. Cheminf. 7, S5

[19] Manning, C. D. et al. (2014) The Stanford CoreNLP Natural Language Processing Toolkit in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60

[20] Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

[21] Swain, M. C. and Cole, J. M. (2016) ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. J. Chem. Inf. Model. 56, 1894– 1904

[22] Cooper, C. B. et al. (2019) Design-to-device Approach Affords Panchromatic Co-sensitized Solar Cell. Adv. Energy Mater. 9, 1802820

[23] Court, C. J. and Cole, J. M. (2018) Auto-generated Materials Database of Curie and Néel Temperatures via Semi-supervised Relationship Extraction. Sci. Data 5, 180111

[24] de Castro, P. B. et al. (2020) Machine-learning-guided discovery of the gigantic magnetocaloric effect in HoB2 near the hydrogen liquefaction temperature. NPG Asia Mater. 12, 35

[25] Beard, E. J. et al. (2019) Comparative Dataset of Experimental and Computational Attributes of UV/vis Absorption Spectra. Sci. Data 6, 307

[26] Mukaddem, K. T. et al. (2019) ImageDataExtractor: A Tool to Extract and Quantify Data from Microscopy Images. J. Chem. Inf. Model. 60, 2492–2509

[27] Hiszpanski, A. M. et al. (2020) Nanomaterials Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge. J. Chem. Inf. Model. 6, 2876–2887

[28] Huang, S. and Cole, J. M. (2020) A database of battery materials auto-generated using ChemDataExtractor. Sci Data 7, 260

[29] Pennisi, E. et al. (2001) The human genome. Science, 291, 1177–1180

[30] Miga, K. H. et al. (2020) Telomere-to-telomere assembly of a complete human X chromosome. Nature 585, 79–84

[31] Dunkel, M. et al. (2006) SuperNatural: a searchable database of available natural compounds. Nucleic Acids Res. 34, D678–D683

[32] Banerjee, P. et al. (2015) Super Natural II—a database of natural products. Nucleic Acids Res. 43, D935–D939

[33] Sorokina, M. and Steinbeck, C. (2020) Review on natural products databases: where to find data in 2020. J. Cheminform. 12, 20

[34] Littlewood, P. B. (2013) Probe the Infinite Variety. Nature 503, 464

[35] Galli, S. (2014) X-ray Crystallography: One Century of Nobel Prizes. J. Chem. Educ. 91, 2009−2012

[36] Saal, J. E. et al. (2013) Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). JOM 65, 1501-1509

[37] Kirklin, S. et al. (2015) The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. npj Comput. Mater.s 1, 15010

[38] Curtarolo, S. (2012) AFLOW: An automatic framework for high-throughput materials discovery. Comput. Mater. Sci. 58, 218-226

[39] Draxl, C. and Scheffler, M. (2018) NOMAD: The FAIR concept for big data-driven materials science. MRS Bull. 43, 676–682

[40] Zagorac, D. et al. (2019) Recent developments in the Inorganic Crystal Structure Database:

theoretical crystal structure data and related features. J. Appl. Cryst. 52, 918–925

[41] Brown, D. G. and Boström, J. (2016) Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? J. Med. Chem. 59, 4443−4458

[42] Ramakrishnan, R. et al. (2014) Quantum chemistry structures and properties of 134 kilo molecules. Sci. Data 1, 140022

[43] Frade, A. P. et al. (2020) Increasing the performance, trustworthiness and practical value of machine learning models: a case study predicting hydrogen bond network dimensionalities from molecular diagrams. CrystEngComm 22, 7186-7192

[44] Bryant, M. J. et al. (2018) Predicting mechanical properties of crystalline materials through topological analysis. CrystEngComm 20, 2698–2704

[45] Devogelaer, J.-J. et al. (2020) Co-crystal Prediction by Artificial Neural Networks. Angew. Chem., Int. Ed. 59, 2–10

[46] Olsthoorn, B. et al. (2019) Band Gap Prediction for Large Organic Crystal Structures with Machine Learning. Adv. Quantum Technol. 2, 1900023

[47] Mazhnik, E. and Oganov, A. R. (2020) Application of machine learning methods for predicting new superhard materials. J. Appl. Phys. 128, 075102

[48] Mazhnik, E. and Oganov, A. R. (2019) A model of hardness and fracture toughness of solids. J. Appl. Phys. 126, 125109

[49] Goodfellow, I. et al. (2014) Generative Adversarial Nets. In Advances in Neural Information Processing Systems 27; (Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., Weinberger, K. Q., Eds.); pp 2672−2680, Curran Associates, Inc..

[50] Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. arXiv (Machine Learning (stat.ML)), 2013, arxiv:1312.6114. https://arxiv.org/abs/1312.6114

[51] Xie, T. and Grossman, J. C. (2018) Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. Phys. Rev. Lett. 120, 145301

[52] Court, C. J. et al. (2020) 3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning. J. Chem. Inf. Model. 60, 4518−4535

[53] 2019 U.S. Department of Energy Basic Energy Sciences Roundtable on Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning, DOI: 10.2172/1630823 https://science.osti.gov/-/media/bes/pdf/reports/2020/AI-ML_Report.pdf

Highlights

Data-driven materials discovery is becoming possible thanks to the rise in artificial intelligence, high-performance computing, and open-data government regulations that have together formed a ‘perfect storm’ for 'big data' analytics.

The success of data-driven materials discovery for a given field of research is often contingent on having available a large and diverse a set of chemical data which display patterns according to structure-property relationships that are associated with that field.

Manually-curated chemical databases have largely grown according to the need for a) technological innovation, b) understanding life, c) characterizing chemical structures, d) synthetic chemistry.

‘Chemistry-aware’ natural language processing tools can auto-generate materials databases for a bespoke application, using the scientific literature as its data source.

Machine-learning methods can be applied to experimental or computational data, or both, in order to classify or optimize such patterns in data.

A design-to-device approach to materials discovery aims to drive down the average 20 year ‘molecule-to-market’ timeframe by which new materials can be realized. This drive has been motivated by the Materials Genome Initiative.

Outstanding Questions Box

What is the precise number of possible chemicals that could exist in the Universe?

When will the metrology of single compounds be replaced by a systems approach to data-driven materials discovery?

How will the shape of chemical data change once data-driven materials discovery becomes mainstream?

How will the balance of computationally and experimentally-generated data shift once data-driven materials discovery becomes mainstream?

How can natural-language processing methods be improved so that push-button materials databases can be instantly dialled-up to meet the needs of machine learning?

How can machine-learning methods engage optimally with a given size and shape of a chemical dataset to produce the best material predictions?

How can uncertainty quantification become sufficiently integrated into machine-learning methods so that material predictions are considered to be entirely trustworthy?

Glossary

Bayesian Optimization: A form of statistical inference that optimizes a model using a probabilistic approach that is subject to prior conditions of certain variables.

ChemDataExtractor: a ‘chemistry-aware’ natural language processing tool that mines text from scientific documents and automatically collates this text into the form of a chemical database.

Chemical space: The set of chemicals that exist within a scope of a given data-science problem.

Generative Adversarial Networks (GANs): A machine-learning method that is composed of a training set of data, a generative neural network (NN) and a discriminative NN. The generative NN takes input data and generates samples from latent distributions, much in the same way that a VAE operates (see definition below). Each generated data sample is compared against a training data set using the discriminative NN that determines if the sample is similar or different to the training data.

Gradient-boosting tree algorithm: An algorithm that performs a machine-learning method which samples a set of trees from a hierarchical model of questions (a decision tree), determines the residual error for each of them and generates an optimized model that minimizes these errors for the decision tree so that its learning ability is ‘boosted’.

Materials Genome Initiative: A white paper that was launched by The White House in 2011 to stimulate science policy, resources and infrastructure in the United States to help discover and exploit new materials for technological innovation much faster and cheaper than the current average 20 year ‘molecule-to-market’ timeframe.

Natural Language Processing (NLP): a sequence of computer-programming operations that enable text to be mined from documents.

Artificial Neural Network: A machine-learning method that comprises a layer of input nodes, a layer of output notes and n layers of hidden nodes that lie inbetween, when presented in its simplest form. Known data are used to train the neural network by optimizing weights on each node such that the input and output data correspond optimally to the known data. The network model thus created then receives new (unseen) input data and subjects them to these weightings to produce predicted output. A neural network with many hidden layers is said to be capable of ‘deep learning’.

Support Vector Machine (SVM): A machine-learning (ML) method that classifies or optimizes a model using regression-based data analysis. It is a supervised ML method, meaning that it needs to train a model with a full set of known input and output parameters of that model, before it can be used to infer model predictions using new (unseen) data.

Variational Auto-Encoders (VAEs): A machine-learning (ML) method that possesses a probability distribution model of discrete parameters that are stored in a latent space which is bounded by two neural networks: an encoder and decoder. Known data are used to create the probability distribution models and the entire model architecture is trained using this latent space. New input data then sample the latent space to generate new information (from the decoder) based on statistical inference (from the encoder). It is known as a generative machine-learning method since it can generate entirely new information.

Figure captions

Figure 1. The data conveyor belt that creates and processes experimental data to which data science methods are applied to afford a 'molecule-to-market' pipeline. Raw data from an experiment are processed, interpreted scientifically, and then reported in the scientific literature. These data can be mined and collated with similar data to form a materials database, using ‘chemistry-aware’ natural language processing software tools. Data analytics can probe such a database to predict new materials chemistry. These predictions are validated by experiment, the results of which are fedback into the experimental lab to inform the nature of the next set of raw data that is created.

Figure 2. The growth of chemical space driven by: a) technological need, b) life, c) chemical structure, d) synthetic reactions.

Figure 3. Various types of chemical data and machine-learning methods that are combined to afford predictive models for data-driven materials discovery. Each example application illustrates a corresponding text description that appears in the paper.

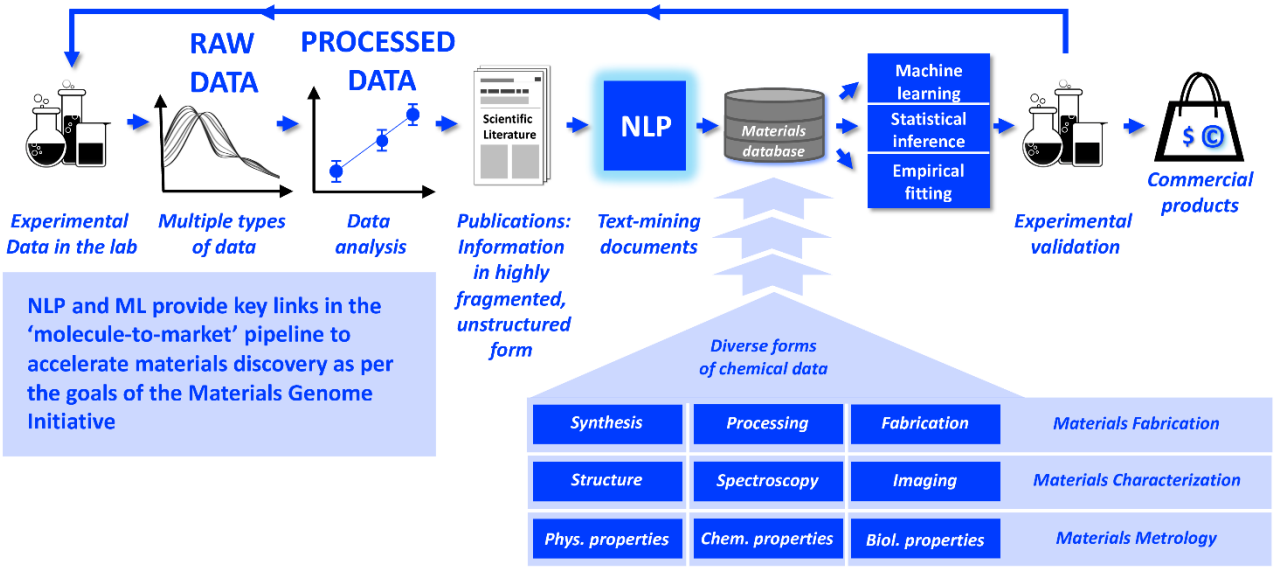
RAW DATA
PROCESSED DATA
Scientific Literature
NLP
Materials database
Machine learning
Statistical inference
Empirical fitting
$ ©
Commercial products
Experimental Data in the lab
Multiple types of data
Data analysis
Publications: Information in highly fragmented, unstructured form
Text-mining documents
Experimental validation
NLP and ML provide key links in the 'molecule-to-market' pipeline to accelerate materials discovery as per the goals of the Materials Genome Initiative
Diverse forms of chemical data
Synthesis
Processing
Fabrication
Materials Fabrication
Structure
Spectroscopy
Imaging
Materials Characterization
Phys. properties
Chem. properties
Biol. properties
Materials Metrology

Figure 1.

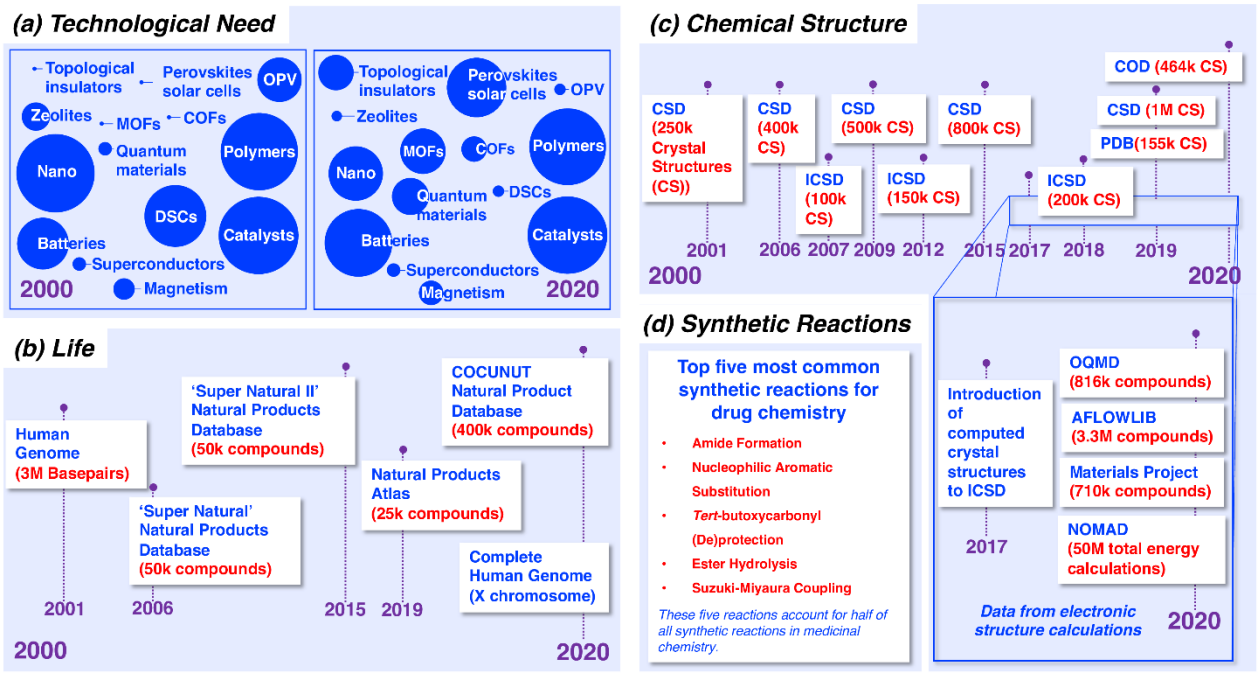
(a) Technological Need
Topological insulators
Perovskites solar cells
OPV
Zeolites
MOFs
COFs
Nano
Quantum materials
Polymers
DSCs
Catalysts
Batteries
Superconductors
Magnetism
2000
2020
(b) Life
Human Genome (3M Basepairs)
'Super Natural' Natural Products Database (50k compounds)
'Super Natural II' Natural Products Database (50k compounds)
Natural Products Atlas (25k compounds)
COCUNUT Natural Product Database (400k compounds)
Complete Human Genome (X chromosome)
2001
2006
2015
2019
2000
2020
(c) Chemical Structure
CSD (250k Crystal Structures (CS))
CSD (400k CS)
ICSD (100k CS)
CSD (500k CS)
ICSD (150k CS)
CSD (800k CS)
ICSD (200k CS)
COD (464k CS)
CSD (1M CS)
PDB(155k CS)
2001
2006
2007
2009
2012
2015
2017
2018
2019
2000
2020
(d) Synthetic Reactions
Top five most common synthetic reactions for drug chemistry
Amide Formation
Nucleophilic Aromatic Substitution
Tert-butoxycarbonyl (De)protection
Ester Hydrolysis
Suzuki-Miyaura Coupling
These five reactions account for half of all synthetic reactions in medicinal chemistry.
Introduction of computed crystal structures to ICSD
2017
OQMD (816k compounds)
AFLOWLIB (3.3M compounds)
Materials Project (710k compounds)
NOMAD (50M total energy calculations)
Data from electronic structure calculations
2020

Figure 2.

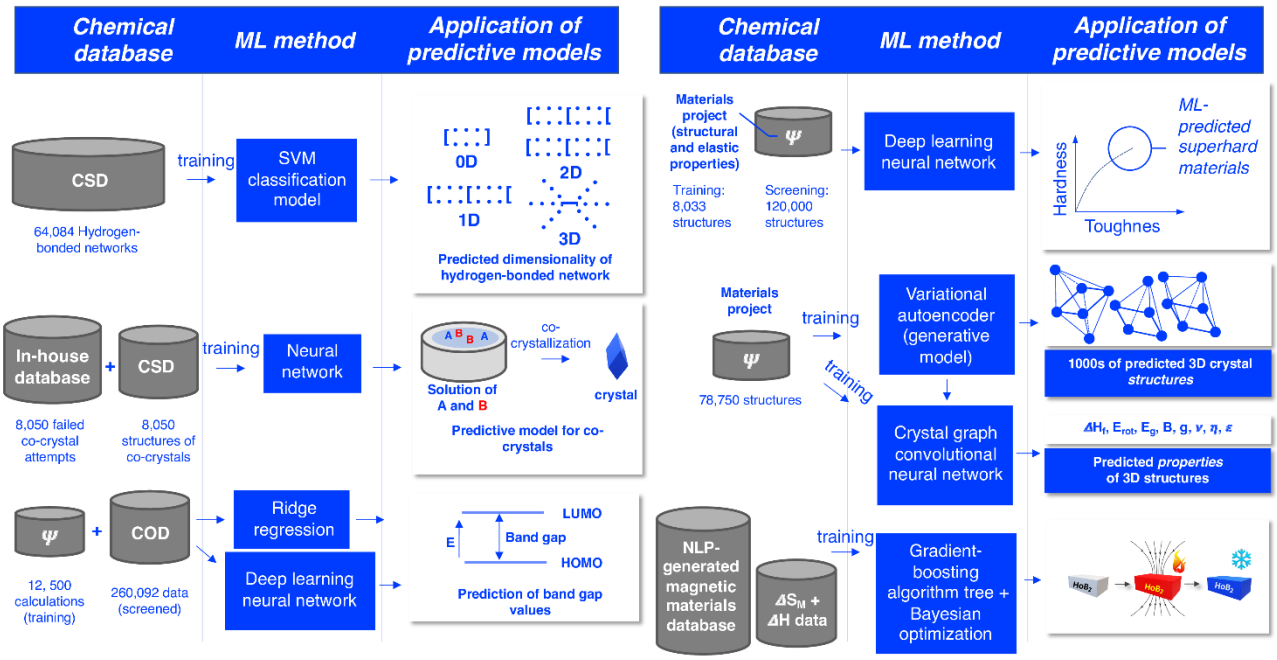
Chemical database
ML method
Application of predictive models
CSD
64,084 Hydrogen-bonded networks
training
SVM classification model
0D
1D
2D
3D
Predicted dimensionality of hydrogen-bonded network
In-house database
8,050 failed co-crystal attempts
CSD
8,050 structures of co-crystals
training
Neural network
co-crystallization
Solution of A and B
crystal
Predictive model for co-crystals
ψ
12, 500 calculations (training)
COD
260,092 data (screened)
Ridge regression
Deep learning neural network
LUMO
Band gap
HOMO
E
Prediction of band gap values
Chemical database
ML method
Application of predictive models
Materials project (structural and elastic properties)
Training: 8,033 structures
Screening: 120,000 structures
Deep learning neural network
Hardness
Toughnes
ML-predicted superhard materials
Materials project
78,750 structures
training
training
Variational autoencoder (generative model)
1000s of predicted 3D crystal structures
Crystal graph convolutional neural network
ΔHf, Erot, Eg, B, g, ν, η, ε
Predicted properties of 3D structures
NLP-generated magnetic materials database
ΔSM + ΔH data
training
Gradient-boosting algorithm tree + Bayesian optimization

Figure 3.