MOLECULAR SIMULATION
2018, VOL. 44, NO. 11, 921–929
https://doi.org/10.1080/08927022.2018.1471692

Taylor & Francis
Taylor & Francis Group

Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space

Johannes Hachmann[a,b,c], Mohammad Atif Faiz Afzal[a], Mojtaba Haghighatlari[a] and Yudhajit Pal[a]

[a]Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, NY, USA; [b]Computational and Data-Enabled Science and Engineering Graduate Program, University at Buffalo, The State University of New York, Buffalo, NY, USA; [c] New York State Center of Excellence in Materials Informatics, Buffalo, NY, USA

ABSTRACT
The use of modern data science has recently emerged as a promising new path to tackling the complex challenges involved in the creation of next-generation chemistry and materials. However, despite the appeal of this potentially transformative development, the chemistry community has yet to incorporate it as a central tool in every-day work. Our research program is designed to enable and advance this emerging research approach. It is centred around the creation of a software ecosystem that brings together physics-based modelling, high-throughput *in silico* screening and data analytics (i.e. the use of machine learning and informatics for the validation, mining and modelling of chemical data). This cyberinfrastructure is devised to offer a comprehensive set of data science techniques and tools as well as a general-purpose scope to make it as versatile and widely applicable as possible. It also emphasises user-friendliness to make it accessible to the community at large. It thus provides the means for the large-scale exploration of chemical space and for a better understanding of the hidden mechanisms that determine the properties of complex chemical systems. Such insights can dramatically accelerate, streamline and ultimately transform the way chemical research is conducted. Aside from serving as a production-level tool, our cyberinfrastructure is also designed to facilitate and assess methodological innovation. Both the software and method development work are driven by concrete molecular design problems, which also allow us to assess the efficacy of the overall cyberinfrastructure.

ARTICLE HISTORY
Received 9 January 2018
Accepted 26 April 2018

KEYWORDS
Chemical space; machine learning; data-driven design; cyberinfrastructure; data science

1. Introduction

The two principal challenges in creating new chemistry and materials are that their behaviour is governed by complicated structure–property and structure–activity relationships [1–3], and that chemical space is practically infinite [4–6]. Traditional experiment-driven trial-and-error approaches are increasingly ill equipped to meet these challenges on their own, in particular since advanced systems require more and more intricate property profiles [7–9]. However, chemical research is currently undergoing a dramatic transformation that is offering new solutions to complex discovery and design problems [10]. After decades of continuous advances in methods, algorithms and computer hardware, the fields of modelling and simulation have reached a tipping point, and they are finally at a stage where they can make accurate predictions for systems that are both realistic and relevant. Progress is now increasingly driven by computational studies, which have become crucial assets in the pursuit of novel chemistry. By making guiding predictions, they can boost the efficiency of research endeavours and uncover promising targets for the more time- and resource-intensive investigations in the laboratory. In addition, they can provide unique insights beyond the scope of empirical observation and thus contribute a solid foundation that underpins new findings. Still, the usual focus on individual compounds has so far been limiting the utility of computational research. While there is obvious value in characterising particular systems of interest, the insights gained in these small-scale studies cannot easily be transferred or generalised.

The shift towards a data-driven discovery and rational design paradigm (cf. Figure 1) promises to mitigate many of the inefficiencies and shortcomings that are still prevalent in contemporary chemical research. There is now a growing agreement on the value of incorporating modern data science – the 4th pillar of science – into chemical research, and this development has been recognised by high-profile funding programs such as the White House Materials Genome Initiative [11]. Yet, despite impressive pioneering efforts (e.g. [12–16]), there is still a distinct disconnect between the promise of this approach and the realities of every-day research in the chemistry community, where data-driven work does not yet play a significant role.

In Section 2 of this review, we will introduce our analysis of this situation and identify objectives as well as associated challenges that form the focal point of our research program. Section 3 provides an outline and rationale of the approach we pursue, followed by a discussion of our cyberinfrastructure work (Section 4), methodological advances (Section 5) and application projects (Section 6). Our discussion is summarised in Section 7.

2. Overarching research objectives and challenges

We have identified three key obstacles that need to be overcome before data science can grow into a mainstay of the chemistry

CONTACT Johannes Hachmann hachmann@buffalo.edu

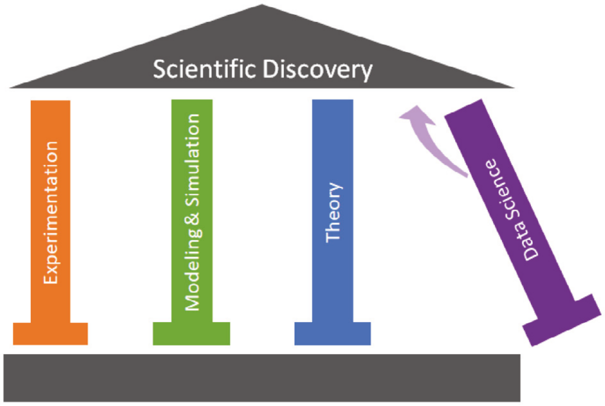


Figure 1. (Colour online) The rise of data science as the 4th pillar of science.

community: (i) data-driven research is practically beyond the scope and reach of most chemists due to a lack of available and accessible tools – the existing tools and expertise tend to be in-house, specialised, or otherwise unavailable to the community at large; (ii) many fundamental and practical questions on how to make data science work for chemical research remain unresolved; (iii) data science is not part of the formal training of chemists, and the community thus oftentimes lacks the necessary experience and expertise to utilise it. Our research program is designed to address these challenges.

The notion to utilise data science in the chemical context is so recent that much of the basic infrastructure has not yet been developed or is still in its infancy [10,17]. Our work seeks to fill this infrastructure gap by creating and deploying an open, general-purpose software ecosystem that fuses computational modelling, virtual high-throughput screening and big data analytics. The former facilitates the large-scale *in silico* exploration of chemical space. Its uncharted domains are expected to hold new classes of compounds and reactions with exceptional properties. We employ machine learning and informatics to mine the resulting data-sets in order to develop an understanding of the hidden structure–property relationships that govern the behaviour of complex chemical systems. These insights are a prerequisite for hyperscreening, rational design and inverse engineering capabilities [18–21]. A key consideration is to make this cyberinfrastructure as comprehensive, integrated, black-box and user-friendly as possible, so that it can readily be employed by interested researchers without the need for excessive expert knowledge. With this approach, we aim to replicate the transformative impact that the emergence of computational chemistry program packages with similar traits have had on the role of modelling and simulation in chemical research today. In addition to delivering a production-level tool for the community, our software ecosystem also serves as a development platform and testbed for innovation in the underlying methods, algorithms, protocols and workflows. The central question this work tackles is how to transfer the success of data science in other domains to problem settings in chemistry. Our work emphasises the value of a general-purpose perspective over focusing on specific solutions for individual application problems that only a relatively small number of experts find interesting and that only provide anecdotal evidence towards our overarching research questions. The development of our tools is nonetheless guided by concrete, real-life molecular design problems, which also allow us to assess the efficacy of our overall approach.

3. Outline and rationale of our research approach

We have developed a basic template for data-driven *in silico* research, which addresses the inherent challenges in the discovery and design of new chemistry, in particular as part of integrated joint ventures with experimentalists. It provides the foundation and framework for our research program, and its rationale can be summarised as following:

- Using computational modelling and simulations, we can rapidly and efficiently assess the properties, behaviour, and performance potential of candidate compounds, materials, and/or chemical transformations for a given problem setting [22–24].
- By combining modelling and simulation with high-throughput screening techniques, we can characterise candidates on a massive scale. These studies naturally lead to big data scenarios [25–29].
- Using modern database technology, we can readily store and access the resulting data sets, e.g. to identify candidates with desired property combinations for on-demand applications [30,31].
- In addition to the immediate information obtained for these thousands or even millions of candidates, we can mine the generated data in its entirety. Using machine learning, we can gain insights into the mechanisms that determine their characteristics and cast these findings into predictive models [32–34].
- By identifying these design rules as well as high-value moieties, building blocks, structural patterns or more general features, we can accomplish the *de novo* design of next-generation candidates [35–37]. Using the predictive models, we can conduct hyperscreenings, i.e. screenings based on data-derived models that typically surpass the scale of the original screenings (based on physics-derived models) by several orders of magnitude.
- Experimentalist partners can pursue the top candidates from the (hyper-)screening and/or *de novo* design. This guidance allows the experimentalists to focus on highly promising targets and avoid wasted efforts on unpromising ones [38]. Additional in-depth modelling and simulations can contribute further insights to the experimental findings, which allow for the advanced optimisation of lead candidates.
- The experimental results can be included as training data in the machine learning approaches. In addition, they can be used to validate, benchmark, calibrate and potentially improve the physics-based modelling and simulation protocols, which closes the design loop [39].

4. Cyberinfrastructure development

We have identified four components (detailed in Sections 4.1–4.4) as critical to realising this approach and unlocking its full potential. Our cyberinfrastructure work pursues the concerted development of four corresponding program suites – *ChemLG*

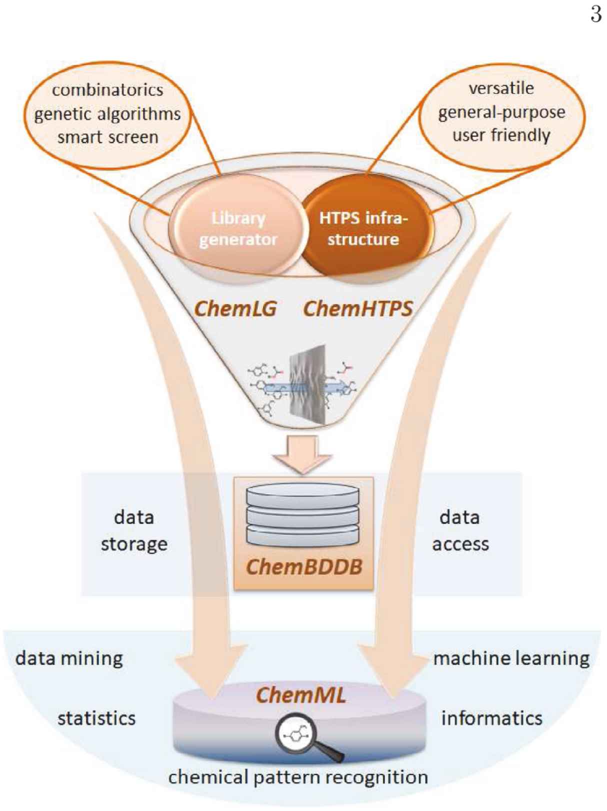


Figure 2. (Colour online) Schematic and connectivity of the software ecosystem comprised of the *ChemLG*, *ChemHTPS*, *ChemBDDB*, and *ChemML* codes.

[40], *ChemHTPS* [41], *ChemBDDB* [42] and *ChemML* [43] – that constitute the software ecosystem outlined in Figure 2 and detailed in Section 4.5.

4.1. Screening library generator

A prerequisite for the high-throughput exploration of chemical space is access to suitable, large-scale screening libraries. We have thus been developing *ChemLG*, a generator for compound and material candidate libraries as well as reaction networks. The key for a successful library generation approach is to balance the ambition for a systematic and exhaustive enumeration of the combinatorial search space, with the need for an efficient and responsive scheme. The generation of combinatorially exhaustive libraries is relatively straightforward, but it rapidly becomes impractical for screening purposes due to its exponential growth. For instance, the largest small-molecule library, GDB-17, contains 166.4 billion molecules, which were generated using only up to 17 atoms (C, N, O, S and halogens) [44]. Most currently available codes are thus limited to generating small, drug-like molecules.

We find that instead of exploring a limited chemical domain exhaustively, it is often more useful to bias the search in directions where candidates are most promising and synthetically viable. In our development of *ChemLG*, we have thus augmented the combinatorial schemes by a number of 'smart' modules that make use of additional input. To address the concern that virtual candidates may not be accessible or desirable (e.g. from a synthetic perspective), we have introduced a constrained-growth scheme that continually prunes the generation process. It accepts user-defined constraints, e.g. to exclude certain structural patterns, building block combinations or sequences; to limit size or chemical makeup; or to enforce symmetries or other rules (e.g. Lipinski's rule). A more advanced version will employ pattern matching and fingerprint similarity cutoffs trained against libraries of known compounds. We can also employ an on-the-fly prescreening through rapid candidate assessment *via* data-derived prediction models (cf. Section 4.4). These models can also serve as fitness functions in *ChemLG*'s genetic algorithm module, which allows us to optimise candidate pools and the chemical structures they contain for specific target applications. *ChemLG*'s 'smart' modules can thus facilitate self-regulating growth of the candidate libraries and ultimately a self-optimising traversal of chemical space. It offers options for the various approaches mentioned above to provide the most suitable solution for different problem settings. We have successfully employed *ChemLG* to produce screening libraries for a number of application projects as shown in Section 6.

4.2. Virtual high-throughput screening infrastructure

A prerequisite for conducting computational high-throughput screening studies is a software infrastructure that can facilitate the execution of thousands or even millions of modelling calculations. For this task, we have been developing the *ChemHTPS* program suite. It designed to streamline and automatise the setup of project environments and directory structures, the generation of job pools based on user-defined candidate libraries (e.g. from *ChemLG*) and modelling protocols, their submission to available hardware in an orderly and load-balanced fashion, job monitoring, error handling, as well as the parsing and bookkeeping of returning results. These processes follow generalised workflow templates that we have been developing from abstraction in the course of different application projects. Key concerns in the development of *ChemHTPS* have been flexibility, reusability and standardisation, as well as the need to accommodate a variety of research fields, modelling and simulation engines, as well as hardware environments. *ChemHTPS* provides the necessary interfaces to *ChemLG*, local queuing systems, compute engines, project databases, and primary data archives. It currently supports the ORCA [45], Q-Chem [46] and GROMACS [47] modelling packages, and bindings to other quantum chemistry, molecular dynamics and solid state physics codes are planned for the future. Given the required user input for the candidate library and modelling protocols, we can now set up and launch a high-throughput *in silico* screening project like the Clean Energy Project [22,26,27,48–50] from scratch in a few minutes, which is a dramatic reduction from its original lead time of several months. We have been using *ChemHTPS* in a number of studies as detailed in Section 6.

4.3. Database infrastructure

The use of modern database technology is of particular importance in the context of data-intensive research. Despite their great utility and despite being essential for projects that accumulate large data-sets, databases are still rarely featured in chemical research. We have been developing the *ChemBDDB* code to simplify and streamline the use of databases and thus

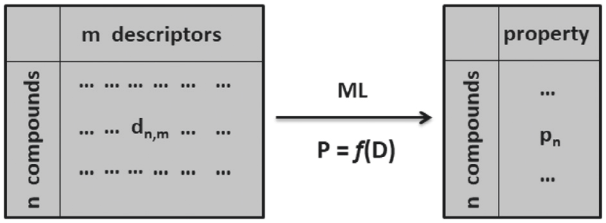


Figure 3. Structure–property relationship mapping *via* machine learning: The structures of the n compounds of a given data-set are represented in an m-dimensional descriptor space, resulting in an $n \times m$-dimensional descriptor matrix D. This descriptor matrix is to be connected to the known properties of the given compounds, which are written as the property vector P of length n. We are seeking the complex mapping function f that encapsulates the structure–property relationship $P = f(D)$ by means of machine learning.

make them more accessible to non-expert users in the chemistry community. *ChemBDDB* provides an automated database setup, a data model template that can readily be customised and the necessary tools to access and manipulate the database. As in *ChemHTPS*, we have been developing the corresponding workflows by abstracting our experience from real-life application projects with flexibility and reusability in mind.

4.4. Data analysis, mining and modelling infrastructure

We have been developing the *ChemML* program suite to establish data analysis, mining and modelling capabilities that allow us to apply state-of-the-art machine learning and informatics methodology to chemical and materials data-sets. *ChemML*'s principal tasks are the creation of predictive regression models and chemical pattern recognition/classification [51]. The former can be thought of as complex mappings between input data (i.e. features of chemical systems) and output observations (i.e. properties) as shown in Figure 3.

The resulting mapping functions (i.e. data-derived prediction models) are usually much easier to evaluate than physical models (e.g. the Schrödinger equation), so using them allows us to dramatically accelerate the characterisation of chemical systems and it thus enables the hyperscreening of chemical space. A typical *ChemML* workflow encompasses a number of distinct steps that can be categorised in six main tasks: (1) input/output, (2) preprocessing, (3) learning, (4) validation, (5) evaluation and (6) visualisation (summarised in Figure 4).

ChemML provides facilities for all these tasks *via* classes of methods. These are either accessed from advanced third-party libraries and stand-alone programs (e.g. scikit-learn [52], Tensorflow [53], Keras [54], Dragon [55], openBabel [56], RDKit [57]), if these represent the state of the art for certain tasks, or from *ChemML Library* where we compile our new/original contributions as well as existing methods that are not otherwise accessible. The feature representation and the machine learning approach are two particularly important aspects in the generation of data-derived models, and they determine a model's predictive performance. *ChemML* can readily call many standard machine learning (as well as preprocessing and visualisation) methods. Examples of available algorithms include multivariate regression [58], support vector machines [59], artificial neural networks [16,60] and deep learning [61]. To support chemistry applications, *ChemML* interfaces the core learning algorithms with domain-specific tools, such as the feature space of molecular descriptors [62,63], topological fingerprints [64,65] or more recent developments such as the Coulomb matrix [32] and bag-of-bonds [66] descriptors. These feature spaces are the abstract 'basis set' in terms of which a machine learning model for a particular structure–property relationship is numerically expressed [67], and *ChemML* provides a comprehensive collection of existing schemes. While our current work focuses on supervised learning, we plan to broaden our scope to unsupervised and reinforcement learning in the future.

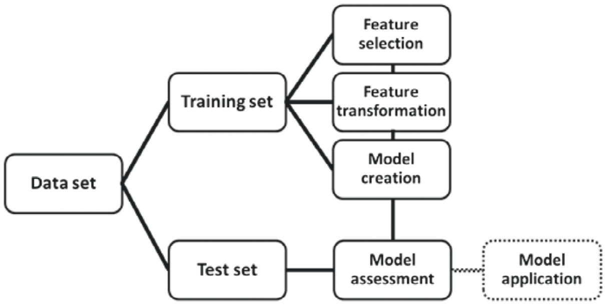


Figure 4. Prototypical workflow for the creation of a data-derived prediction model in *ChemML*: A given data-set is split up into a training and a test set. The data in the training set is expressed in a feature space and subject to feature selection and potentially feature transformation. Once a suitable representation is found, we can generate the desired prediction model as shown in Figure 3. We evaluate the predictive performance of the model by applying it to the independent test set and assess the resulting prediction error. (The assessment process, e.g. *via* k-fold cross-validation, is typically more involved than shown in this simplified schematic.) If the error is acceptable, we can use this model to make predictions for compounds outside the given data-set.

4.5. Software ecosystem

The four program packages discussed in the previous sections (*ChemLG*, *ChemHTPS*, *ChemBDDB*and *ChemML*) are loosely connected, i.e. they can either be employed as a comprehensive unit (see Figure 2), or in combination with drop-in replacements (e.g. with a different library generator or a custom database engine), or as standalone applications. The development work on our software ecosystem includes conceptual work, the design and assessment of protocols and workflows, the formulation of guidelines and best practices and the implementation of both glue code and new methods (cf. Section 5). The key cyberinfrastructure challenges include the robust abstraction of workflows for general-purpose applications, scaling issues (e.g. associated with expensive data generation and the combinatorial nature of chemical space), code sustainability, as well as platform and distribution issues. While we emphasise black-box automation to reach non-expert users, we allow full customisation of all settings (in particular in *ChemML*). We continually extend and refine the features and capabilities of all four codes. These improvements are driven by feedback from application projects both inside and outside our group (cf. Section 6). This input from different real-world application problems is a key to making this cyberinfrastructure as resourceful as possible. All codes will be open and freely shared with the community under 3-clause BSD licence.

There are a number of exciting, high-profile software development efforts along similar lines by others in the field (e.g. [28, 68]). However, despite great popular demand, there is currently no cyberinfrastructure for data-driven *in silico* research that is accessible to the community, applicable to a wide range of chemical problems, and that offers a level of comprehensiveness, automation and integration comparable to that of modern computational chemistry program packages. Our contribution pursues this niche, which stands out in its scope and prospective utility.

5. Methodological advances

Adapting data science techniques from other application domains for the study of chemical systems requires a substantial rethinking and redevelopment of the existing methodology [69–72]. There are still considerable challenges that need to be overcome to make data-driven research in chemistry a success. Our software ecosystem is designed as a development platform and testbed that allows us to systematically, rapidly and efficiently assess the utility and performance of existing techniques as well as new ones that we are introducing in the course of our work. Its modular structure makes it easy to implement new techniques for any component of the workflow, and this flexibility and versatility allows our cyberinfrastructure to streamline a traditionally arduous standalone process: instead of writing full prototype implementations to test each new idea, we only have to write the specific corresponding modules. Once added to the software ecosystem (e.g. *via ChemML Library*), we immediately have access to the comprehensive suite of tools it provides, including the facilities for the automated testing of these new methods against many existing ones with essentially no overhead in human time. During the benchmarking of, e.g. a new machine learning technique, *ChemML* generates prediction models for a collection of data-sets using the new approach, performs *k*-fold cross-validation and systematically compares its prediction errors against those from established ones. For the assessment of the prediction errors (i.e. with respect to the known results for the test set), *ChemML* provides a full suite of statistical metrics. New techniques that are competitive compared to existing ones or that offer other benefits (e.g. interpretability, efficiency) become part of the release branch of *ChemML Library* and are thus directly accessible to users as part of the larger package. The step from prototype to production implementation is therefore limited to individual modules.

Our ongoing development efforts include the creation of new methods that allow us to (a) tailor the design of screening libraries and efficiently survey the practically infinite chemical space; (b) target the training data generation process to maximise its information yield and minimise its computational cost; (c) streamline the use of databases and advance issues on ontologies and semantics that support knowledge creation; (d) formulate and set up chemical problems for optimised knowledge extraction and incorporate the underlying physics into the creation of suitable feature spaces; (e) pick the ‘right’ machine learning approach and hyperparameters for a given problem setting; (f) identify the sweet spot between over- and underparameterising models (i.e. models with too much vs. too little flexibility); (g) assess the soundness, range of validity and predictive performance of data-derived models beyond the empirical cross-validation of test sets; (h) go from prediction models to inverse engineering and rational design and (i) extract chemical knowledge (i.e. causalities) from the statistical correlations underlying these prediction models.

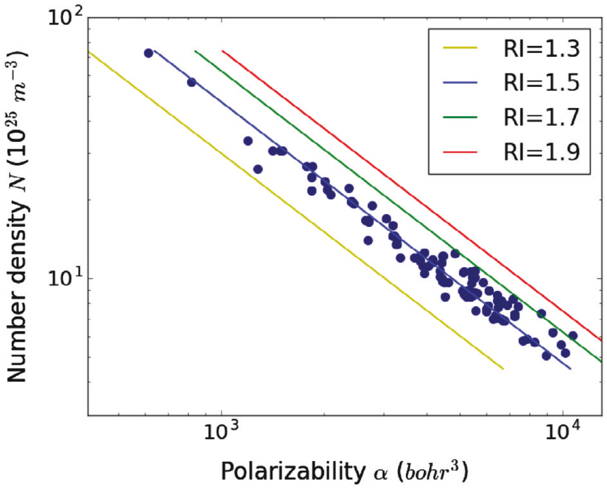


Figure 5. (Colour online) Representation of the refractive index value domains in the polarisability α vs. number density N parameter space with the results from our validation study.

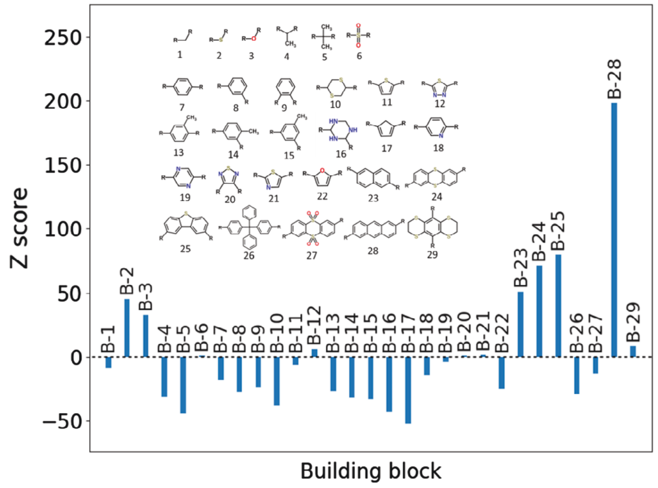
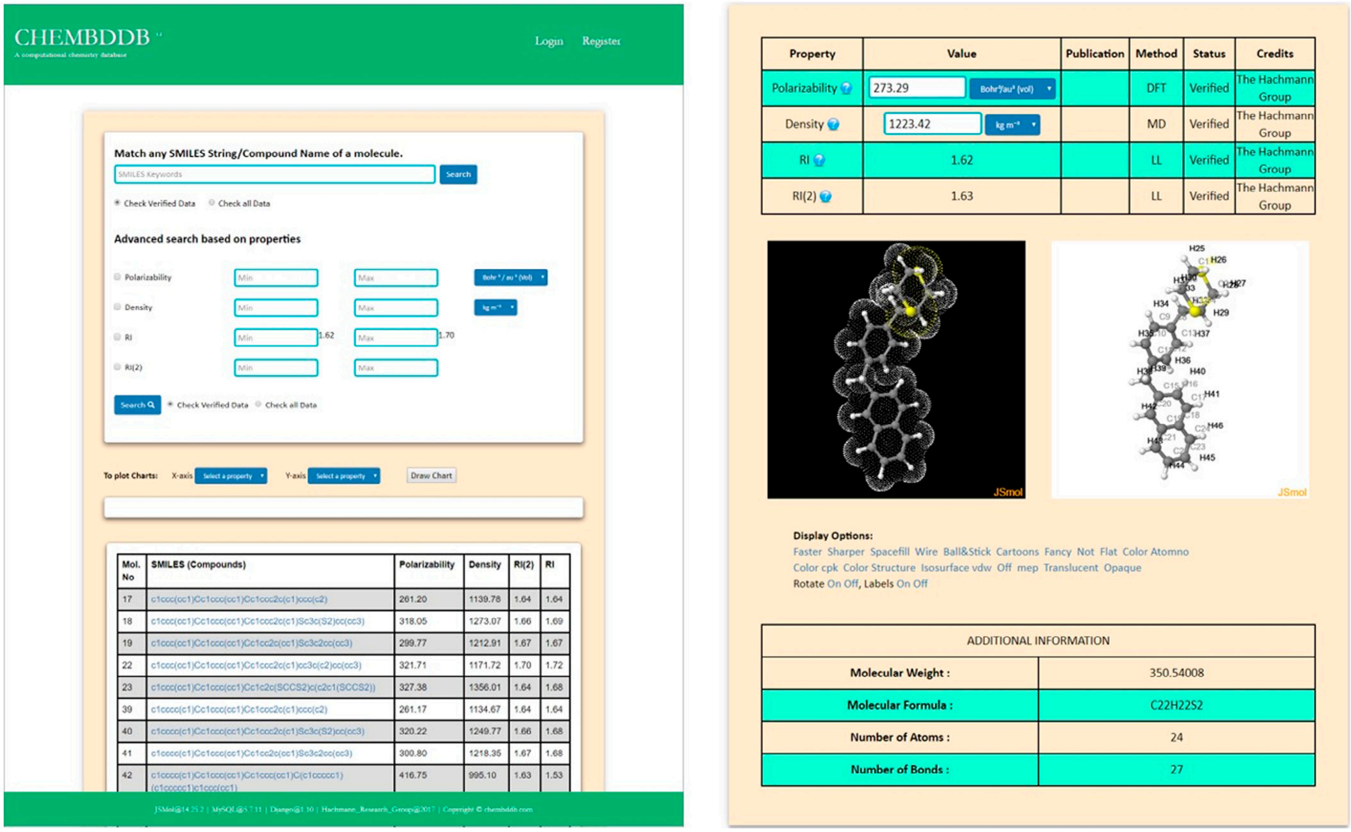


Figure 6. (Colour online) Hypergeometric distribution analysis highlighting the prevalence of certain building blocks in our top candidates compared to the overall screening library.

In addition, we pursue automated machine-learning-approach optimisation and meta machine learning to establish guidelines, black-box features and defaults that provide added value to our cyberinfrastructure. The idea is to automatically generate fully optimised machine learning models and subsequently (machine) learn, which machine learning approach performs best for different classes of chemical problems and types of data-sets. (The machine learning approach refers to the machine learning method, feature representation and hyperparameters. The class of problems typically refers to different classes of compounds/reactions and target properties. The type of data-set refers to the nature of the given data and addresses the ‘4 Vs of big data’, i.e. its volume, veracity, variety and velocity.) The problem we aim to solve is that the use of machine learning is so new in the chemical context and to the chemical community that decisions about the approaches to be employed are often



CHEMBDDB
A computational chemistry database

Login Register

Match any SMILES String/Compound Name of a molecule.

SMILES Keywords [Search]

● Check Verified Data ○ Check all Data

Advanced search based on properties

□ Polarizability	Min	Max	Bohr³ / au³ (Vol)
□ Density	Min	Max	kg m⁻³
□ RI	Min 1.62	Max 1.70	
□ RI(2)	Min	Max	

[Search] ● Check Verified Data ○ Check all Data

To plot Charts: X-axis [Select a property] Y-axis [Select a property] [Draw Chart]

Mol. No	SMILES (Compounds)	Polarizability	Density	RI(2)	RI
17	c1ccc(cc1)Cc1ccc(cc1)Cc1ccc2c(c1)ccc(c2)	261.20	1139.78	1.64	1.64
18	c1ccc(cc1)Cc1ccc(cc1)Cc1ccc2c(c1)Sc3c(S2)cc(cc3)	318.05	1273.07	1.66	1.69
19	c1ccc(cc1)Cc1ccc(cc1)Cc1cc2c(cc1)Sc3c2cc(cc3)	299.77	1212.91	1.67	1.67
22	c1ccc(cc1)Cc1ccc(cc1)Cc1ccc2c(c1)cc3c(c2)cc(cc3)	321.71	1171.72	1.70	1.72
23	c1ccc(cc1)Cc1ccc(cc1)Cc1c2c(SCCS2)c(c2c1(SCCS2))	327.38	1356.01	1.64	1.68
39	c1cccc(c1)Cc1ccc(cc1)Cc1ccc2c(c1)ccc(c2)	261.17	1134.67	1.64	1.64
40	c1cccc(c1)Cc1ccc(cc1)Cc1ccc2c(c1)Sc3c(S2)cc(cc3)	320.22	1249.77	1.66	1.68
41	c1cccc(c1)Cc1ccc(cc1)Cc1cc2c(cc1)Sc3c2cc(cc3)	300.80	1218.35	1.67	1.68
42	c1cccc(c1)Cc1ccc(cc1)Cc1ccc(cc1)C(c1ccccc1)(c1ccccc1)c1ccc(cc1)	416.75	995.10	1.63	1.53

JSMol@14.25.2 | MySQL@5.7.11 | Django@1.10 | Hachmann_Research_Group@2017 | Copyright © chembddb.com

Property	Value	Publication	Method	Status	Credits
Polarizability	273.29 Bohr³/au³ (vol)		DFT	Verified	The Hachmann Group
Density	1223.42 kg m⁻³		MD	Verified	The Hachmann Group
RI	1.62		LL	Verified	The Hachmann Group
RI(2)	1.63		LL	Verified	The Hachmann Group

Display Options:
Faster Sharper Spacefill Wire Ball&Stick Cartoons Fancy Not Flat Color Atomno
Color cpk Color Structure Isosurface vdw Off mep Translucent Opaque
Rotate On Off, Labels On Off

ADDITIONAL INFORMATION	
Molecular Weight :	350.54008
Molecular Formula :	C22H22S2
Number of Atoms :	24
Number of Bonds :	27

Figure 7. (Colour online) Project database for our high-refractive index polymer work.

made *ad hoc* and without deeper consideration, guidance or insights. The automated optimisation implemented in *ChemML* offers both a brute-force grid search and a more advanced genetic algorithm to identify the best machine learning approach (within a given search space) for specific problems at hand that yield the models with the smallest prediction errors. In the meta machine learning work, we compile the findings of these optimisations together with descriptors characterising the corresponding data-sets and chemical problems, and mine these data to systematically derive insights and technical guidelines as mentioned before. We are incorporating the latter into our cyberinfrastructure to enforce best practices and prevent misuse/pitfalls (e.g. overfitting and flawed validation). These outcomes will advance the utility and reliability of machine learning in chemistry and boost our knowledge and understanding of its performance and inner workings. These insights also help us focus the machine-learning-approach optimisation and save computing time.

Examples of our work include the development and implementation of a new trend-based feature selection (TBFS) method in molecular descriptor space [73]. For proof-of-principle, we performed a study on a density functional theory data-set of 1.8 million compounds from the Clean Energy Project [22,26,27,48–50], which took about 5000 CPU years to generate. Our TBFS, in combination with a simple ridge regression model, reproduced the principal energy levels of this data-set in just 15 minutes with a mean absolute error of 0.07 eV (1.3%). In other work, we used pattern matching to study the inheritance of certain building block properties in copolymers [73,74] and derived a projection model between different quantum chemical approximations that allows us to obtain high-level results from cheap, low-level calculations for an array of properties [73]. We also generated a random forest classification to predict the exceptions for which the projection fails [73,75].

6. Application projects

Our cyberinfrastructure is designed to enable the study of science driver applications and in particular provide a framework for collaborative research efforts that combine experimental, computational and data mining thrusts.

Several ongoing projects are already using it, including searches for new high-refractive index polymers for optical applications, deep eutectic solvents for supercapacitors and battery electrolytes, molecular hydrolysis catalysts for solar water splitting and fuel cells, doped and defect nanographene anode materials for lithium ion batteries, polyvinyl-based biodegradable polymers for biomedical plastics, liquid organic hydrogen carriers for the hydrogen economy and organic semiconductors for photovoltaics and other applications [22,26,27,48,50,73–77].

For this review, we will outline our work on high-refractive index polymers as a prototypical example of an application study [73,76]. High-refractive index polymers have attracted interest due to their potential application in optic and optoelectronic devices [78] (e.g. image sensors [79,80], displays [81] and light sources [82], in which they can be introduced *in situ* as microlenses [83], waveguides [84], microresonators [85], interferometers [86], anti-reflective coatings [87], optical adhesives [88] and substrates [89]). Most of these applications require a refractive index of 1.7, 1.8, or larger [90], while typical carbon-based polymers only exhibit values in the range of 1.3–1.5 [91]. As the properties of organic polymers can be tailored by controlling their molecular structure [91], they are a prime example for a rational design target. We have employed *ChemLG* to produce suitable screening libraries, e.g. for polyimide-based polymers (270,000 compounds). The refractive index value of organic polymers can be calculated *via* the Lorentz–Lorenz equation, which is parametrised by the polarisability and number density values of a given candidate compound at the polymer limit. We have used *ChemHTPS* to compute the polarisabilities of all candidates (as well as 100,000 small organic molecules as model systems) at the density functional theory level. For the model systems, we also computed the number densities *via* molecular dynamics simulations. Using *ChemML*, we created a deep neural network model of these number density values. With a training set of only 20,000 data points, we achieved an excellent agreement between physics-based simulation and data-derived prediction model with an R^2 of 0.98. We can thus use the deep neural network model as a surrogate for the much more demanding molecular dynamics calculations in order to rapidly compute the number densities needed for the refractive index predictions without significant loss of accuracy. In an alternative route to the number densities, we generated a support vector regression model for the packing factor of amorphous polymers [76]. The validation of its prediction for 112 polymers with experimentally known number densities showed a good agreement with an R^2 of 0.87. Our current working model for the refractive index reproduces the available experimental results with an R^2 of 0.94, and it is thus clearly predictive. Our validation study yielded a mean absolute deviation of 0.010 (0.9%), a root mean square deviation of 0.018 (0.1%) and a maximum deviation of 0.045 or 3.0% with respect to the experimental data [76]. Figure 5 shows its results in the polarisability–number density parameter space.

Using this model, we screened the candidate library discussed earlier as well as a secondary library with 1.5 million compounds. We found multiple candidates with refractive index values higher than the target of 1.8, which we shared with our experimentalist collaborators for further investigation [73]. We subsequently performed a pattern analysis that identified several moieties and moiety combinations that are significantly over-expressed in our top candidates (see Figure 6), and which we thus chose as promising starting points for ongoing design studies [73].

We optimised our machine learning approach and identified, e.g. the most suitable descriptor space for the different target properties. We are also using this application case to develop capabilities for automated hyperparameter and model optimisation. In other ongoing work, we are investigating the utility of purely data-derived prediction models for the refractive index values (and other properties) and contrast them to the hybrid model discussed above, which is built within the framework of the Lorentz–Lorenz equation. Both the *in silico* data as well as experimental data found in the literature are stored in a project database created by *ChemBDDB* (see Figure 7).

7. Conclusions

Our research program recognises the great opportunities that are arising with the shift towards a data-driven *in silico* research paradigm in chemistry, materials science and the corresponding engineering disciplines. It focuses on delivering and deploying a cyberinfrastructure that is filling a distinct infrastructure gap and on building foundations that make data-driven research a viable and widely accessible proposition for the chemistry community. The template for our efforts is the rise of computational chemistry program packages over the past decades and the tremendous impact it has had on the role of modelling and simulation in contemporary chemical research. Following this example, our research program aims to advance our capacity to tackle complex discovery and design challenges, facilitate an increased rate and quality of innovation, improve our understanding of the associated molecular and condensed matter systems and democratise the tools that make these developments possible. We have shown in real-life application projects that this approach indeed offers a path to overcoming some of the prevalent limitations of traditional trial-and-error approaches.

The long-term objective of our work and related efforts by others is to help pioneer a fundamental transformation of the discovery process in chemistry, to make data science an integral part of the chemical enterprise, to shape the transition towards a data-driven discovery and rational design paradigm and to spearhead a broad move by the community along those lines.

Acknowledgements

Computing time on the high-performance computing clusters '*Rush*', '*Alpha*', '*Beta*' and '*Gamma*' was provided by the University at Buffalo (UB) Center for Computational Research (CCR). The work presented here is part of MAFA's, MH's and YP's PhD theses. In addition, this review mentions contributions of the following members and former members of the Hachmann group: Dr. Andrew J. Schultz (Research Assistant Professor; deep eutectic solvent and organic photovoltaics applications), Aditya Sonpal (MSc; *ChemBDDB*, *ChemML*, deep eutectic solvent and high-refractive index polymer applications), Gaurav Vishwakarma (MSc; *ChemML*), Po-Han Chen (MSc; *ChemML*), Vigneshwar Kumaran Sudalayandi Rajeswari (MSc; biodegradable polymer application), Amol Rajendra Mahajan (MSc; liquid organic hydrogen carrier application). Shirish Sivaraj (MSc; *ChemBDDB*), Noah A. Zydel (BSc; *ChemML*), Chi Hin Chan (BSc; deep eutectic solvent application), Andrew J. DeRooy (BSc; deep eutectic solvent application), Sykhere A. Brown (BSc; *ChemML*), Supriya Agrawal (MSc '17; *ChemBDDB*), Sai Prasad Ganesh (BSc '17; high-refractive index polymer application), Mark A. Pitman (BSc '17; hydrolysis catalyst application), William S. Evangelista II (MEng '16; *ChemHTPS*) Yujie Tian (MSc '16; *ChemML*, organic semiconductor application), Mikhail Pechagin (BSc '16; *ChemML*), Ching-Yen Shih (MSc '15; *ChemML*) and Bryan A. Moore (BSc '15; organic photovoltaics application).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work reviewed here has been supported by start-up funds provided through UB, as well as grants by the New York State Center of Excellence in Materials Informatics [1140384-8-75163], [1122381-12-66357], [1114099-1-66357]; the National Science Foundation CBET Energy for Sustainability and CAREER programs [CBET-1511528], [OAC-1751161]; the Army Research Office SBIR program [W15QKN-17-C-0078].

References

[1] Selassie CD. History of quantitative structure-activity relationships. In: Abraham DJ, editor. Burger's medicinal chemistry and drug discovery. Hoboken (NJ): Wiley; 2003. p. 1–48.

[2] Müller K-R, Rätsch G, Sonnenburg S, et al. Classifying 'drug-likeness' with kernel-based learning methods. J Chem Inform Model. 2005;45:249–253.

[3] Le Bailly deTilleghem C, Govaerts B. A review of quantitative structure-activity relationship (QSAR) models. Technical Report 07027, Universite catholique de Louvain; 2007.

[4] Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. Nature. 2004;432:855–861.

[5] Kirkpatrick P, Ellis C. Chemical space. Nature. 2004;432:823.

[6] Dobson CM. Chemical space and biology. Nature. 2004;432:824–828.

[7] Zvinavashe E, Murk AJ, Rietjens IMCM. Promises and pitfalls of quantitative structure-activity relationship approaches for predicting metabolism and toxicity. Chem Res Toxicol. 2008;21:2229–2236.

[8] Scior T, Medina-Franco JL, Do QT, et al. How to recognize and workaround pitfalls in QSAR studies: a critical review. Curr Med Chem. 2009;16:4297–4313.

[9] Schneider G. Virtual screening: an endless staircase? Nat Rev Drug Discovery. 2010;9:273–276.

[10] Rajan K. Informatics for materials science and engineering: data-driven discovery for accelerated experimentation and application. Amsterdam: Butterworth-Heinemann; 2013.

[11] National Science and Technology Council. Materials genome initiative for global competitiveness. Tech. Rep. Washington (DC): National Science and Technology Council; 2011.

[12] Hansen K, Biegler F, Fazli S, et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. J Chem Theory Comput. 2013;9:3404–3419.

[13] Huan TD, Mannodi-Kanakkithodi A, Ramprasad R. Accelerated materials property predictions and design using motif-based fingerprints. Phys Rev B. 2015;92:1–10, 1503.07503v2.

[14] Bartók AP, Kondor R, Csányi G. On representing chemical environments. Phys Rev B. 2013;87:184115.

[15] Isayev O, Fourches D, Muratov EN, et al. Materials cartography: representing and mining materials space using structural and electronic fingerprints. Chem Mater. 2015;27:735–743.

[16] Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett. 2007;98:146401.

[17] Snyder JC, Rupp M, Hansen K, et al. Finding density functionals with machine learning. Phys Rev Lett. 2012;108:253002.

[18] Mueller T, Hautier G, Jain A, et al. Evaluation of tavorite-structured cathode materials for lithium-ion batteries using high-throughput computing. Chem Mater. 2011;23:3854–3862.

[19] Wang S, Wang Z, Setyawan W, et al. Assessing the thermoelectric properties of sintered compounds via high-throughput ab-initio calculations. Phys Rev X. 2011;1:021012.

[20] Wilmer CE, Leaf M, Lee CY, et al. Large-scale screening of hypothetical metal-organic frameworks. Nat Chem. 2012;4:83–89.

[21] Korth M. Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: evaluation of electronic structure theory methods. Phys Chem Chem Phys. 2014;16:7919–7926.

[22] Olivares-Amaya R, Amador-Bedolla C, Hachmann J, et al. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. Energy Environ Sci. 2011;4:4849–4861.

[23] Wen S, Nanda K, Huang Y, et al. Practical quantum mechanics-based fragment methods for predicting molecular crystal properties. Phys Chem Chem Phys. 2012;14:7578–7590.

[24] Stevanović V, Lany S, Ginley DS, et al. Assessing capability of semiconductors to split water using ionization potentials and electron affinities only. Phys Chem Chem Phys. 2014;16:3706–3714.

[25] Potyrailo R, Rajan K, Stoewe K, et al. Combinatorial and high-throughput screening of materials libraries: review of state of the art. ACS Combin Sci. 2011;13:579–633.

[26] Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, et al. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J Phys Chem Lett. 2011;2:2241–2251.

[27] Hachmann J, Olivares-Amaya R, Jinich A, et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard clean energy project. Energy Environ Sci. 2014;7:698–704.

[28] Gunter D, Cholia S, Jain A, et al. Community accessible datastore of high-throughput calculations: experiences from the materials project. In: 2012 Sc Companion: High Performance Computing, Networking Storage and Analysis. Scc; 2012. p. 1244–1251.

[29] White AA. Big data are shaping the future of materials science. MRS Bull. 2013;38:594–595.

[30] Blum LC, van Deursen R, Reymond JL. Visualisation and subsets of the chemical universe database GDB-13 for virtual screening. J Comput Aided Mol Des. 2011;25:637–647.

[31] Ruddigkeit L, van Deursen R, Blum LC, et al. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inform Model. 2012;52:2864–2875.

[32] Rupp M, Tkatchenko A, Müller KR, et al. Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett. 2012;108(5):058301.

[33] Pilania G, Wang C, Jiang X, et al. Accelerating materials property predictions using machine learning. Sci Rep. 2013;3:2810. DOI:10.1038/srep02810.

[34] Mansbach RA, Ferguson AL. Machine learning of single molecule free energy surfaces and the impact of chemistry and environment upon structure and dynamics. J Chem Phys. 2015;142:105101.

[35] Pegg SCH, Haresco JJ, Kuntz ID. A genetic algorithm for structure-based de novo design. J Comput Aided Mol Des. 2001;15:911–933.

[36] Proschak E, Sander K, Zettl H, et al. From molecular shape to potent bioactive agents II: fragment-based de novo design. ChemMedChem. 2009;4:45–48.

[37] Nakamura M, Hachiya T, Saito Y, et al. An efficient algorithm for de novo predictions of biochemical pathways between chemical compounds. BMC Bioinform. 2012;13:S8.

[38] Sokolov AN, Atahan-Evrenk S, Mondal R, et al. From computational discovery to experimental characterization of a high hole mobility organic crystal. Nat Commun. 2011;2:437–438.

[39] Hartenfeller M, Schneider G. De Novo drug design. In: Bajorath J, editor. Chemoinformatics and computational chemical biology. Vol. 672; Totowa, NJ: Humana Press; 2011. p. 299–323.

[40] Hachmann J, Afzal MAF. ChemLG 0.5 – a library generator code for the enumeration of chemical and materials space; 2017. Available from: https://hachmannlab.github.io/chemlg.

[41] Hachmann J, Evangelista WS, Afzal MAF, et al. ChemHTPS 0.7 – an automated virtual high-throughput screening program suite for chemical and materials data generation; 2017. Available from: https://hachmannlab.github.io/chemhtps.

[42] Hachmann J, Agrawal S, Sonpal A, et al. ChemBDDB 0.2 – a big data database toolkit for chemical and materials data storage; 2017. Available from: https://hachmannlab.github.io/chembddb.

[43] Hachmann J, Haghighatlari M. ChemML 0.10 – a machine learning and informatics program suite for chemical and materials data mining; 2017. Available from: https://hachmannlab.github.io/chemml.

[44] Reymond JL, Ruddigkeit L, Blum L, et al. The enumeration of chemical space. Wiley Interdisciplinary Rev Comput Mol Sci. 2012;2:717–733.

[45] Neese F. The ORCA program system. Wiley Interdisciplinary Rev Comput Mol Sci. 2012;2:73–78.

[46] Shao Y, Gan Z, Epifanovsky E, et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. Mol Phys. 2015;113:184–215.

[47] Abraham MJ, Murtola T, Schulz R, et al. Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1–2:19–25.

[48] Amador-Bedolla C, Olivares-Amaya R, Hachmann J, et al. Organic photovoltaics. In: Rajan K, editor. Informatics for materials science and engineering: data-driven discovery for accelerated experimen-

tation and application. Amsterdam: Butterworth-Heinemann; 2013. Chapter 17; p. 423–442.
[49] Pyzer-Knapp EO, Suh C, Gómez-Bombarelli R, et al. What is high-throughput virtual screening? A perspective from organic materials discovery. Ann Rev Mater Res. 2015;45:195–216.
[50] Lopez SA, Pyzer-Knapp EO, Simm GN, et al. The Harvard organic photovoltaic dataset. Sci Data. 2016;3:160086.
[51] Kowalski BR, Bender CF. Pattern recognition. Powerful approach to interpreting chemical data. J Am Chem Soc. 1972;94:5632–5639.
[52] Pedregosa F, Weiss R, Brucher M. Scikit-learn : machine learning in python. J Mach Learn Res. 2011;12:2825–2830.
[53] Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous systems; 2015, software available from tensorflow.org.
[54] Chollet F, et al. Keras; 2015, software available from: https://github.com/fchollet/keras.
[55] Talete srl. DRAGON (Software for Molecular Descriptor Calculation); 2011, software available from: http://www.talete.mi.it/.
[56] O'Boyle NM, Banck M, James CA, et al. Open babel: an open chemical toolbox. J Cheminform. 2011;3:33. DOI:10.1186/1758-2946-3-33
[57] RDKit: Open-source cheminformatics, software available from: http://www.rdkit.org.
[58] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.
[59] Müller KR, Mika S, Rätsch G, et al. An introduction to kernel-based learning algorithms. IEEE Trans Neural Netw. 2001;12:181–201.
[60] Manzhos S, Carrington T. A random-sampling high dimensional model representation neural network for building potential energy surfaces. J Chem Phys. 2006;125:084109.
[61] Dahl GE. Deep learning approaches to problems in speech recognition, computational chemistry, and natural language text processing [PhD thesis]. University of Toronto; 2015.
[62] Todeschini R, Consonni V, Mannhold R, et al. Handbook of molecular descriptors. Weinheim: Wiley-VCH; 2000.
[63] Sykora VJ, Leahy DE. Chemical Descriptors Library (CDL): a generic, open source software library for chemical informatics. J Chem Inform Model. 2008;48(10):1931–1942.
[64] Nilakantan R, Bauman N, Dixon JS, et al. Topological torsion: a new molecular descriptor for sar applications comparison with other descriptors. J Chem Inform Comput Sci. 1987;27:82–85.
[65] O'Boyle NM, Sayle RA. Comparing structural fingerprints using a literature-based similarity benchmark. J Cheminform. 2016;8:1–14.
[66] Hansen K, Biegler F, Ramakrishnan R, et al. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. J Phys Chem Lett. 2015;6:2326–2331.
[67] Ramakrishnan R, von Lilienfeld OA. Machine learning, quantum chemistry, and chemical space. In: Parrill AL, Lipkowitz KB, editors. Reviews in computational chemistry. Vol. 30. Hoboken, NJ: John Wiley &amp;amp; Sons; 2017. p. 225–256.
[68] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput Mater. 2016;2:16028.
[69] Schütt KT, Arbabzadah F, Chmiela S, et al. Quantum-chemical insights from deep tensor neural networks. Nat Commun. 2017;8:13890.
[70] Isayev O, Oses C, Toher C, et al. Universal fragment descriptors for predicting electronic properties of inorganic crystals. Nat Commun. 2017;8:15679.
[71] Ferré G, Haut T, Barros K. Learning molecular energies using localized graph kernels. J Chem Phys. 2017;146:114107.
[72] Collins CR, Gordon GJ, von Lilienfeld OA, et al. Constant size molecular descriptors for use with machine learning. arXiv 2017, arXiv:1701.06649.
[73] Collins CR, Gordon GJ, von Lilienfeld OA, et al. Constant size molecular descriptors for accurate machine learning models of molecular properties. J Chem Phys. 2018;148:241718. DOI:10.1063/1.5020441
[74] Tian Y. Inheritance of molecular orbital energies from monomer building blocks to larger copolymers in organic semiconductors [master's thesis]. University at Buffalo; 2016.
[75] Shih CY. Systematic trends in results from different density functional theory models [master's thesis]. University at Buffalo; 2015.
[76] Afzal MAF, Cheng C, Hachmann J. Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers. J Chem Phys. 2018;148:241712.
[77] Kumaran Sudalayandi Rajeswari V. First-principles modeling of polymer degradation kinetics and virtual high-throughput screening of candidates for biodegradable polymers [master's thesis]. University at Buffalo; 2018.
[78] Lei T, Wang JY, Pei J. Roles of flexible chains in organic semiconducting materials. Chem Mater. 2014;26:594–603.
[79] Angione MD, Pilolli R, Cotrone S, et al. Carbon based materials for electronic bio-sensing. Mater Today. 2011;14:424–433.
[80] Voigt A, Ostrzinski U, Pfeiffer K, et al. New inks for the direct drop-on-demand fabrication of polymer lenses. Microelectron Eng. 2011;88:2174–2179.
[81] Ummartyotin S, Juntaro J, Sain M, et al. Development of transparent bacterial cellulose nanocomposite film as substrate for flexible organic light emitting diode (OLED) display. Indus Crops Prod. 2012;35:92–97.
[82] Xiang C, Ma R. Devices to increase OLED output coupling efficiency with a high refractive index substrate. US Patent 9,640,781. 2017.
[83] Nishiyama H, Nishii J, Mizoshiri M, et al. Microlens arrays of high-refractive-index glass fabricated by femtosecond laser lithography. Appl Surf Sci. 2009;255:9750–9753.
[84] Kokubun Y, Funato N, Takizawa M. Athermal waveguides for temperature-independent lightwave devices. IEEE Photon Technol Lett. 1993;5:1297–1300.
[85] Wei H, Krishnaswamy S. Direct laser writing polymer micro-resonators for refractive index sensors. IEEE Photon Technol Lett. 2016;28:2819–2822.
[86] Rodríguez A, Vitrant G, Chollet PA, et al. Optical control of an integrated interferometer using a photochromic polymer. Appl Phys Lett. 2001;79:461–463.
[87] Singaravalu S, Mayo DC, Park HK, et al. Anti-reflective polymer-nanocomposite coatings fabricated by RIR-MAPLE. In: SPIE LASE. Vol. 8607. International Society for Optics and Photonics; Bellingham, WA; 2013. p. 860718.
[88] Kim JB, Lee JH, Moon CK, et al. Highly enhanced light extraction from surface plasmonic loss minimized organic light-emitting diodes. Adv Mater. 2013;25:3571–3577.
[89] Kim E, Cho H, Kim K, et al. A facile route to efficient, low-cost flexible organic light-emitting diodes: utilizing the high refractive index and built-in scattering properties of industrial-grade PEN substrates. Adv Mater. 2015;27:1624–1631.
[90] Jintoku H, Ihara H. The simplest method for fabrication of high refractive index polymer-metal oxide hybrids based on a soap-free process. Chem Commun. 2014;50:10611–10614.
[91] Liu JG, Ueda M. High refractive index polymers: fundamental research and practical applications. J Mater Chem. 2009;19:8907–8919.