nature portfolio

https://doi.org/10.1038/s42254-022-00518-3

Supplementary information

On scientific understanding with artificial intelligence

In the format provided by the authors and unedited

On scientific understanding with artificial intelligence
Full collected anecdotes

Mario Krenn,[1,2,3,4,*] Robert Pollice,[2,3] Si Yue Guo,[2] Matteo Aldeghi,[2,3,4] Alba Cervera-Lierta,[2,3] Pascal Friederich,[2,3,5] Gabriel dos Passos Gomes,[2,3] Florian Häse,[2,3,4,6] Adrian Jinich,[7] AkshatKumar Nigam,[2,3] Zhenpeng Yao,[2,8,9,10] and Alán Aspuru-Guzik[2,3,4,11,†]

[1] *Max Planck Institute for the Science of Light (MPL), Erlangen, Germany.*
[2] *Chemical Physics Theory Group, Department of Chemistry, University of Toronto, Canada.*
[3] *Department of Computer Science, University of Toronto, Canada.*
[4] *Vector Institute for Artificial Intelligence, Toronto, Canada.*
[5] *Institute of Nanotechnology, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany.*
[6] *Department of Chemistry and Chemical Biology, Harvard University, Cambridge, USA.*
[7] *Division of Infectious Diseases, Weill Department of Medicine, Weill-Cornell Medical College, New York, USA.*
[8] *Center of Hydrogen Science, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China.*
[9] *The State Key Laboratory of Metal Matrix Composites, School of Materials Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China.*
[10] *Innovation Center for Future Materials, Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, 429 Zhangheng Road, Shanghai 201203, China.*
[11] *Canadian Institute for Advanced Research (CIFAR) Lebovic Fellow, Toronto, Canada.*

(Dated: July 8, 2022)

I. INTRODUCTION

Here we collect all responses received by computational physicists, chemists and biologists, answering the following question (received from March-October 2020):

> Dear colleagues,
>
> This is a short/quick request!
>
> My colleague Mario Krenn and I are carrying out a short survey of the actual scientific discovery process for a perspective paper we are writing. I am writing to you because I consider you one of the leaders and/or pioneers of computational science and/or big data/AI/high throughput insight.
>
> In particular, we are interested in the chronological aspects of your discovery/ies with regards to computer-driven insight.
>
> The question is: What is your own best example of the following:
>
> * You look at the output of a computational science simulation or big data / AI / high-throughput simulation exercise and you see an outlier, extremal point, or interesting trend and then it suggests to you a new physical phenomenon which you then actually actively investigate and *understand*. This then leads hopefully to a high-quality result. You could potentially use this understanding now as a *tool without the computation*, ie. you learned a design principle or scientific "law". We are looking for the most "general" outcomes, ie. after you understand your results, you found a generalizable idea.
>
> You could send us more than one example if you want. Please send the reference to the discovery too if you don't mind.
>
> Finally,
>
> * Are you OK with quoting your email and listing your correspondence with us in the paper? We may not select all your responses depending on what we are looking for.
>
> If you could send us by October 10, this would be very useful. Any examples sent to us after that, may not be considered for the paper.
>
> Thank you for your time. We are asking because it is often not clear in papers how you came across the discovery, so the narrative is crucial to our argument.
>
> Alan and Mario

The full responses helped us understand a comprehensive picture of how scientists get a new understanding from computers. The answers contain many excellent accounts which go far beyond what we were able to present directly in the main text. We thank all participants for their time and their significant contributions. Below we ex

* mario.krenn@mpl.mpg.de
† alan@aspuru.com

II. INSIGHTS BY ANASTASSIA ALEXANDROVA

My insight started not form an outlier in a simulation, but from the fact I came to the field of heterogeneous catalysis from the field of chemical physics of small clusters, which gave me a clear idea of what can happen if an ultracold cluster beam gets a few Kelvin too warm: the spectrum would become essentially a mess. Small metal clusters supported on surfaces of semiconductors are used as catalysts, which can be superb, and their properties are very sensitive to atoms count, the nature of the support, and the conditions at which they are used. Modeling in the field of catalysis was always done on a single cluster shape, the global minimum in the best-case scenario. Basically, I could not believe my eyes, because temperatures of 700 K for a small cluster mean thermal access to very many minima on the free energy surfaces, not one. So the catalyst has to contain tens or hundreds of distinct structures, forming a dynamic ensemble in reaction conditions. From there, as any chemist would, I began suspecting that less stable structures (metastable states) of the catalyst could be more reactive, and maybe even responsible for the entire catalysis. So we developed efficient sampling techniques,[1] got in close collaboration with several experimentalists, and started collecting the evidence. We soon found that all observable properties of dynamic catalytic interfaces (activity,[2] selectivity,[3–8] stability,[9, 10] operando spectra[11]) are best described as ensemble-averages, and that the ensemble of catalyst states constantly reorganizes as part of the reaction coordinate. We are still in the process of finding out all of what that means for the theory of catalysis. We see some old things holding true, such as the linearity of the Arrhenius plot.[12] We also see several rules breaking down, such as the scaling relations.[13] We see some new phenomena, such as the impossibility to suppress Ostwald ripening through size-selection. The most general outcome for the moment is just that the catalyst is an ensemble of many states, not one, that the less stable but still accessible members in this ensemble can be more catalytically active, and that the reaction mechanism is not one but a swarm of many.[14] The obvious complication then is finding those true active sites that constantly come and go in reaction conditions. By the way, these sites are not easily pinned down by experiment, because even operando measurements give an ensemble-averaged signal, which is overwhelmed by majority species. Hence, theory is the only player in town that can do it.

III. INSIGHTS BY ROMMIE AMARO

Well, this is the case for quite a few application-oriented studies coming out of the group.

For example, the case with the SARS-CoV-2 spike, which was just published today [15].

We noticed in the simulation of the open structure, that a couple of glycans had some different behaviors in the open/closed conformation of the spike.

We then explored the basis for this unexpected finding and (after a lot of work) learned (predicated first, then confirmed through experiment) that some glycans on the spike do more than just shield it, but they actually act in a coordinated way with the spike protein to participate in the opening mechanism. From this we know now to look for glycans to participate in such phenomena. Opens the door to a lot of new studies / understandings, regarding exploring the role of glycans in biological systems.

Then there was the older example with p53 [16] where we found a pocket in simulations that had not been seen in experiment, saw it could accommodate small molecules, identified molecules that would bind into that site and control the activity of the protein. This is now a quite well-trodden framework for the discovery of cryptic pockets from molecular simulation. Lots of folks trying to figure out what is the 'magic sauce' for elucidation of such sites. Lots and lots of folks. Less clear in this area, what makes a druggable pocket, and what does not.

IV. INSIGHTS BY CURTIS BERLINGUETTE

We sought to optimize the hole mobility of a doped and annealed organic semiconductor using our self-driving laboratory, Ada. We configured Ada to autonomously explore a wider range of doping levels than we would typically study with manual experiments. This autonomous search led to an unexpected scientific finding: Ada's ML-driven optimization revealed a region of enhanced thermal stability at high doping levels.[17] This result was surprising because high dopant levels typically reduce the thermal stability of these types of films.[18–20] With these new results, we have drawn the conclusion that dopants with a higher intrinsic thermal stability can govern the thermal stability of glassy, organic semiconductors. Leveraging dopants to stabilize organic semiconductors is a compelling concept that we continue to explore. The use of machine vision, in tandem with autonomous experiments, played a key role in leading us to this finding (see recent preprint entitled Quantifying defects in thin films using machine vision[21]).

V. INSIGHTS BY LILLIAN CHONG

My story is about how my lab got into force field development as a result of a "failed" simulation that we ran on the Anton special-purpose supercomputer in 2011. Instead of abandoning this failure, a closer

examination of the results led to an entire PhD thesis for my graduate student, Karl Debiec, with three publications. This work has also opened up a new NSF-funded research direction in my lab.

The failed simulation involved the sticking together of the domains in a two-domain protein after running for 1 microsecond and therefore would not have been detected by simulations on typical computing resources. This result contradicted NMR experiments, which revealed that the domains tumble independently of each other in solution. Upon further analysis, we discovered that the reason why the domains were sticking together was because the domains were interacting via salt bridges (or pairs of hydrogen bonded, oppositely charged amino acids) and these salt bridges were being overstabilized by the simulation model (force field). In fact, as we demonstrated in the attached paper, nearly all of the force fields at the time shared this same issue of overstabilizing salt bridges. Interestingly, the force fields that yielded more reasonable salt bridge propensities involved atomic charges with implicit solvent polarization and one of these force fields was the Amber ff14ipq force field developed by David Case [22].

We proceeded to collaborate with David to develop an entirely new force field (Amber ff15ipq) that addresses the salt bridge issue. This force field demonstrates the power of using first principles to develop simulation model, exhibiting not only reasonable salt bridge propensities, but (i) the expected balance of secondary structures for both globular and non-globular ("disordered") peptides/proteins, and (ii) good agreement with NMR observables, including J-coupling constants that are just as accurate (if not more) than force fields that were specifically parameterized to reproduce the experimental values. Amber ff15ipq has been available in both the widely-used Amber and OpenMM software packages.

VI. INSIGHT BY GERARDO CISNEROS

Our example has to do with discovery and characterization of cancer mutations and their impact on the structure and function of specific proteins. The initial question arose from computational simulations of the reaction mechanism of a DNA repair enzyme (DNA polymerase lambda, PolL). Our calculations yielded predictions on catalytically important residues that had not been investigated and are conserved among many human polymerases. This led to the question: Are there any DNA mutations related to some cancer that can result in changes in these residues? To answer this question we developed a protocol to analyze large-scale genomic data from multiple genome-wide association studies (GWAS) to uncover single nucleotide polymorphisms (SNPs) associated with a particular phenotype (disease), statistical analysis to determine the significance of the SNPs, followed by computational modeling to determine the effects of the mutation. The software for genomic data mining is agnostic to phenotype but we have used it for cancer databases. The original paper is [23]. One of our first predicted cancer mutations, which associates PolL with breast cancer, was subsequently confirmed experimentally [24]. Since then, we've developed a database for cancer mutations on DNA repair and modification proteins [25]. and have used our approach to investigate several systems [26–28].

In the above cases, we used our method to discover the cancer mutation and predict the effect of the resulting protein variant with computational analysis, followed by experimental confirmation by collaborators.

VII. INSIGHT BY ANDY COOPER AND GRAEME DAY

A. Andy Cooper

I think our strongest example so far is our work on crystal structure prediction with Graeme Day (cc'd here; [29] and (particularly) [30] – Energy Structure Function maps). The primary physical phenomenon here (so far) has been porosity. The idea behind CSP and ESF maps is that you can predict the most likely crystal packing (and therefore function) of molecules. By definition, these packings are hard to predict, and hence doesn't exactly lead to "a tool without computation" in a general sense. However, by working with enough of these systems and studying the structure landscapes in detail (and we've now looked at lots), you can begin to build up some kind of intuitive feeling for the likely crystal packing of a new molecule without doing the computation. (Intuition in this sense will be less reliable than computational prediction, but it is faster.)

The reason that this is valuable is that one structure landscape reveals a wide array of possible crystal packings with their associated lattice energies – it is a little bit like looking at the 'structural DNA' for the molecule. By contrast from a practical standpoint, most molecules have between one and a handful of polymorphs; as such, you need to work on a lot of systems over a career to gain similar insights.

In a nutshell, by working with CSP / ESF maps, there is the possibility to gain insights into structure-property relationships at a much faster rate than working with practical systems alone. There are some interesting questions about how to codify this – to give one specific example, we just published an example [31] where it was possible to map out the likelihood of a candidate molecule exhibiting pi-pi stacking, which is important in organic electronics and photocatalysis.

In precis: ESF maps can be thought of as a guide

to intuitive crystal structure design (although philosophically, I lean toward using these things as quantitative predictions, rather than a "tool without computation").

Graeme can probably add to this.

B. Graeme Day

I could add a couple of things to what Andy has already said.

First is to support the idea that landscapes of predicted structures help develop our intuition about the relationships between molecular structure and crystal packing faster than through studying observed crystal structures. As Andy already said, we get access to the structure-stability relationships in large sets of crystal structures. Through ESF maps, we also learn about structure-property relationships. In the simplest case, we are looking at a set of crystal structures all of one molecule, so structure-stability-property relationships are unobscured by changes in molecular properties, which we normally have to deal with when looking to the crystallographic databases or other collections of observed crystal structures. Where we have looked at how changes in molecular structure affect crystal packing, CSP (and ESF maps) lets us see how the entire landscape of possible structures is influenced and, so, how certain intermolecular interactions and molecular packing motifs can be strongly structure-directing. This feeds into the design process of what molecules to look at next. So, there is some element of developing "tools without computation".

An example of where we have learnt from observing outliers in large structure sets is the area of porous molecular crystals. In the study described in one of the papers that Andy mentioned[30] - we (luckily) looked at the crystal energy landscapes over a very wide energy range. The field of crystal structure prediction has, up to this point, focussed very heavily on getting the ranking of crystal structures correct at and around the global lattice energy minimum. Very rarely have people looked further than 10-20 kJ/mol above the global minimum. When looking at the CSP results for these molecules, we visualised the landscapes up to about 100 kJ/mol from the global minimum. It was this view that showed us 'spikes' on the energy vs density view of the CSP landscapes. The spikes were sets of predicted structures that dropped down from the main energy/density distribution: structures that were much more stable than they should be for their packing density. Although a plot of energy vs density is a simplified picture of a high dimensional energy surface, we had the idea that these spikes correspond to isolated, deep regions of lattice energy - structures that would a high energy barrier hindering conversion to a denser structure. These would have been missed completely if we had only looked at the low energy region.

These spikes were outliers in that original study, but are now a key feature that we look for when testing ideas for formation of porous molecular crystals. When we see these features on CSP landscapes / ESF maps, we now have confidence that we can create the corresponding structures in the lab. Initial reports on CSP results often now start with "Spikes!" or "Sadly, no spikes" as a more important outcome than the structure of the global energy minimum. Predicting these energy landscape features has recently motivated further experimental work on trimesic acid [32] and the photocatalyst study that Andy has already mentioned [31].

VIII. INSIGHT BY FRANÇOIS-XAVIER COUDERT

Although it's definitely a subjective question, one of the most useful "general principle" idea derived from screening large number of materials in my group's research has been when the findings run contrary to conventional wisdom. One such example is in the area of mechanical properties and meta materials: if you look at the existing literature, there are quite a few studies were the highlight is the experimental measurement or computational calculation of a negative Poisson's ratio in some crystallographic direction or other. This has been published in a number of zeolites and metal-organic frameworks, usually accompanied by the broad statements that this behaviour is counter-intuitive (which is subjective, but most people would agree) and rare (which is never backed up by references).

Using computational tools and screening of 13.621 inorganic compounds[33], we wanted actually found that this behaviour is not rare at all. What is very rare (0.3% of the crystals studied) is the occurence of complete auxeticity, i.e. negative Poisson's ratio in all directions of space. But the existence of some direction of negative Poisson's ratio is actually common, occurring in 30% of structures. I think this definitely changed our view (and hopefully our colleagues' too!) of that particular property, beyond the method of calculation and the details, to what we expect for this property.

Another example is to try and use databases and/or machine learning approaches to learn more about the "hidden laws" behind certain materials. We've been looking for some time at zeolites, where the question of experimental feasibility is still not fully answered: out of so many possible zeolitic frameworks, all relatively low in energy, why is it that only a small number are experimentally accessible? There again, looking at mechanical properties hints at a possible component to that answer[34, 35] where the mechanical properties of the framework are playing a role, in addition

to thermodynamic criteria. This is part of an answer in the formation mechanism of these zeolites, an eminently complex phenomenon (high T, in solution with a lot of species, etc.) that could not be obtained without systematic analysis of a significant database.

IX. INSIGHT BY LEE CRONIN

Controlling an organic synthesis robot with machine learning to search for new reactivity[36] – Outline: The manuscript describes a high-throughput closed-loop robot searching organic reactivity using on-line analytics such as NMR, IR, and MS. The feedback loop allowed for efficient searching through reactivity space of organic molecules. New discoveries found can be classified as outliers, for example, high molecular weight peaks in mass spectra with m.w. ¿ 500 were indicators of multicomponent reactions. Hits from the robot were actively identified and investigated yielding several new transformations.

A curious formulation robot enables the discovery of a novel protocell behaviour[37] – Outline: The manuscript describes the exploration of droplet-based protocells using a closed-loop robotic platform equipped with Curiosity Algorithm. The Curiosity Algorithm explores self-propelling multicomponent oil formulations and their emerging behaviours in an open-ended way with no specified target in the observational space. The paper demonstrated much faster exploration of droplet behaviours as compared to random parameter search in the defined observational phase space. Utilizing this closed-loop system, we were able to discover extreme response of droplet behaviours with minute temperature changes. We selected novel experimental behaviours from the explored phase space and performed in-depth mechanistic studies to investigate the relationship between dynamic behaviour of droplets and input chemical composition.

Evolution of oil droplets in a chemorobotic platform[38] – Outline: In this project we conducted our research from two different starting points or hypotheses: 1) Life is the product of evolution and natural selection. Therefore our experiments were guided by a Genetic Algorithm (GA) which mimicked how natural selection works. GAs have been used successfully on many different applications, and therefore and our research question was: Can a GA be used to evolve chemistry into life-like behaviours? 2) Droplets have been widely used to model protocells. We conducted a exhaustive literature search focusing on droplets of a few millimeters of size (thus, no microfluidic experiments) that expressed a life-like behaviour, such as movement, division or chemotaxis. Once we had a few good candidates, we manually tested them on a Petri dish, and we selected the droplet recipes that were more active and the ones that had divergent chemical properties, such as different densities, solubilities,... Once our ingredients were chosen, the first experimental step was to generate a high-throughput grid search across all the ingredients to test that their combinations could produce interesting behaviours. The results were qualitative assessed, and once we decided that the recipe-space was interesting, we decided to run the GA for three different user-defined fitness functions: movement, division and vibration. In total, around 20 000 one-minute experiments were generated. These experiments were analysed using Support Vector Regression with a Radial Basis Function Kernel, and we produced a series of fitness landscapes that directly map from droplet recipes to droplet behaviour. With this fitness landscapes we could point to any point and generate a desired behaviour without having to go through the GA again.

X. INSIGHT BY ELISA FADDA

As a computational biophysicist specialized in HPC-based biomolecular simulations, I am quite used to the analysis of large-sized (from a few Gb to several Tb) noisy datasets.

More specifically, I study the structure, dynamics and molecular recognition of complex carbohydrates (or glycans) and how that relates to their many different biological functions. This information is extremely hard to get experimentally because glycans are among the most intrinsically disordered biomolecules in life. HPC has enormous potentials in advancing glycoscience, yet the molecular dynamics simulations we run reflect precisely the glycans disordered nature, which makes the rather chaotic time (or energy)-evolution data and "ball of yarn"-looking structures, quite difficult to disentangle.

Despite their 'structural disorder' common denominator, glycans sequences are not random, but follow precise rules. This always suggested to me that not all glycans may be equally disordered and that their sequence and branching holds the key to the broad diversity of their functional roles. Determining how does these sequence-to-structure-to-function links work out is one to the main research topics in my lab.

Our analytical protocol consists in running statistical analysis (via clustering, KDE, PCA) on the conformational sampling data collected for every single joint (or glycosidic linkage) that connects the monosaccharide units in the often-large tree-like glycan structures. This largely reduces the complexity of the data and indicates us trends and outliers that we usually check at this point by visually analysing the time-evolution of the structures we produce, or trajectories.

By analysing the "joints" data in the context of their immediate environment, i.e. the types of monosaccharides present and how are they linked, we have been able to find quite interesting patterns

repeating[39, 40], i.e. the whole glycan 3D structure and dynamics can be rationalized in terms of "glycoblocks"[1], i.e. groups of 2 to 3 monosaccharides with precise branching, that differ by shape and flexibility. Different glycan architectures expose different glycoblocks to receptors for recognition, supporting many experimental findings that very (deceptively) similar glycans sequences have very different functions and can be highly immunogenic[39].

This rather simple glycoblock architecture can be extremely useful not only to understand sequence-to-structure-to-function relationships in complex carbohydrates, but also to inform the design of synthetic carbohydrate sequences with a desired structure to advance functional and recognition studies. In my lab we are now developing methods for the automated reconstruction of glycans structures from glycoblocks, which will bypass the need of computationally expensive simulations and the expert user input.

XI. INSIGHT BY RAFAEL GOMEZ-BOMBARELLI

In my independent career, the best highlight is the story in this paper: ML as a tool to gather enough robust data from experiments, a graph-metric to be able to measure distances and cluster, and materials chemistry insight arising from the analysis of the clusters[41].

XII. INSIGHT BY LETICIA GONZALEZ

Photophysics of Thionucleobases – In "The origin of effcient triplet state population in sulfur-substituted nucleobases" [42] we discuss the origin of the very special photophysical behaviour of the thionucleobase compound class. Thionucleobases are closely related to the biologically important nucleobases that form the genetic code inside DNA. The only difference between nucleobases and thionucleobases is the fact that the latter have at least one oxygen atom replaced by the homologous sulfur atom. These two classes of compounds generally behave very similarly—thionucleobases can even replace nucleobases in DNA and RNA. However, they have very different photochemistry and photophysics, where nucleobases relax in an ultrafast manner back to the ground state, whereas thionucleobases form long-lived triplet states through intersystem crossing very fast and with almost unit yield.

Our initial observation was based on multiple experimental papers [43–49] that gave a very consistent picture that all investigated thionucleobases—2-thiouracil, 2-thiothymine, 2-thiocytosine, 4- thiothymine, 2,4-dithiothymine, 6-thioguanine—showed very consistent photophysical behavior. In all these

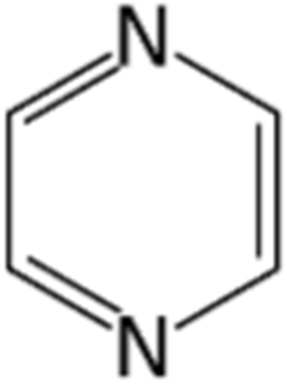

Figure 1. Pyrazine

bases, intersystem crossing populates triplet states on a few-picosecond time scale with very high yield, leading to long-lived excited states. Another observation was that all these thionucleobases show a strongly red-shifted absorption spectrum compared to their parent nucleobases. Stimulated by these findings, several computational studies[50–56] strived to explain the photophysical behavior of the thionucleobases, although each study focused on only a single compound (chronologically: 6-thioguanine, 4-thiothymine, 2-thiouracil). In Ref. [42], after studying 2-thiocytosine and comparing to the other bases, we noticed some commonality in the potential energy surfaces of the thionucleobases. Thus, we attempted to find an in-depth explanation for their photophysics. We optimized the relevant excited-state minima and excited-state–ground-state crossing points for 6-thioguanine, 2-thiouracil, 2-thiocytosine, and their respective parent nucleobases. The results showed clearly that the introduction of a sulfur atom into a carbonyl group leads to strongly stabilized excited-state minima but does not stabilize the crossing points that enable the characteristic short excited-state lifetimes of the canonical nucleobases. Essentially, the photophysics of the nucleobases that is dominated by the aromatic ring is overwritten by the photophysics of a thiocarbonyl group because the latter gives rise to the lowest-energy excited states.

XIII. INSIGHT BY JOHANNES HACHMANN

This is not from my own research, but the attached Guardian clip about the work by Hod Lipson was really what got me hooked on ML. I came across this back when I was with Alan at Harvard and it convinced me that ML has tremendous potential in the physical and chemical sciences. I think this is still one of the coolest and most impressive uses of machine learning in our neck of the woods, in particular the presented analysis of the motion tracking data for the

double pendulum, which is a prototype of a chaotic system with exponentially diverging trajectories. This clip is also one of the first things I show my students when I introduce the idea of using data science[57].

XIV. INSIGHT BY ROALD HOFFMANN

A story of numbers to ideas – Early on, in the brief three years when extended Hückel calculations could be called state-of-the art, I computed the orbitals of pyridine and the three diazines. You can tell how early in my career that was – I had no coworkers [58]. At that time, the delocalization of the MOs was by itself a publishable result. For the 1,4-diazabenzene, pyrazine [Figure 1], two molecular orbitals mainly localized on N, were seen. Two striking things about these N-lone pairs, (for that is what they roughly looked like – in those days there were no easy contour-plotting programs; one looked at the coefficients), emerged: they were split in energy by several eV, and the antisymmetric combination of the two was lower in energy.

Both facts provided some numerical satisfaction in comparison with experimental reality, as they explained some hitherto known but not understood spectroscopy of pyrazine. In time, when those experiments became available, Edgar Heilbronner's measurements of the photoelectron spectra of pyrazine confirmed the "anomalous" order of the orbitals, antisymmetric below symmetric. The result remained on the face of it puzzling – if the lone pairs were localized, they should not overlap much directly (the distance between those N was 2.7A), and to the extent they did, the symmetric combination (the bonding one) should be at slightly lower energy.

We (now I had some coworkers, both were undergraduates at Cornell) came up with an explanation, involving the symmetry-conditioned overlap/interaction of the lone pairs with the uniquely disposed CC bond (and its * counterpart) between them, as shown in Figure 2. [59]

The explanation was not only pretty, it could be extended to other molecules (the benzynes, for instance; also to other heterocycles, such as 1,4-diazabicyclo[2.2.2]octane). The tests of the explanation were first numerical – did the effect occur in other molecules with the same disposition of the bonds? How did it depend on rotation around intervening bonds? In time experimental tests, the measurements I mentioned, came in. And the analysis provided a way of thinking about orbital interactions in general – through-space and through-bond.

It was satisfying to have the numbers, to explain the spectra. It was much more satisfying, a joy indeed, to come up with the explanation. And had the initial computational result (the splitting of the energies of the lone-pair combinations) been an artifact of the computational method, then this "explanation" would have joined the junkpile of wrong theories. The numbers, and that they were moderately reliable, were essential. But one did not, must not, stop with the numbers.

Another time I will tell you how the fact that the underlying method was not numerically reliable, actually served to strengthen the qualitative conclusions drawn from it by a human being.

XV. INSIGHT BY JAN HALBORG JENSEN

My work is generally focussed on maximising som property, i.e. finding the outliers.

So far we've had little luck rationalising what we found. For example, in this study[60] we looked for insulating molecules. We were not able to see what the top 5 candidates (Table 1) had in common.

I think we'll see more and more of this. Chemistry is very complex and usually influenced by many competing forces. While each force is understood conceptually, what is needed for a prediction for a complex molecule is a quantitative estimate of each to see what dominates.

Another example, is the prediction of regioselectivity of electrophilic aromatic substitution. Heuristic prediction is trivial for simple molecules (e.g. a single heteroaromatic ring), but for complex molecules (several heteroaromatic rings) there are too many competing factors to consider. This is why we made regiosqm [61]

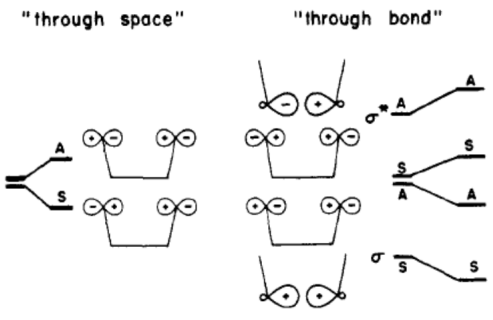

Figure 2. Interaction of two orbitals separated by three bonds. At left the two orbitals might be interacting via a direct, "through-space" overlap. S and A stand for symmetric or antisymmetric combination of lone pairs with respect to a mirror plane interchanging the orbitals. At right the S and A p orbital combinations interact in a predictable way (the orbitals involved "repel" each other) with and * orbitals of the localized bond of the central bond. The net result is an A below S splitting of the middle orbital set.

XVI. ADRIAN JINICH

Note from authors: Adrian Jinich is also a co-author of the manuscript.

Example 1: – In this work we used quantum chemistry to predict the standard redox potentials of 652 biochemical redox reactions [62].

The resulting dataset led us to a deeper understanding of why the redox cofactor NAD(P), with its reduction potential ranging between -370 mV and -250 mV, has the prime role supporting cellular redox reactions in almost all organisms, participating in most (¿50%) known redox reactions. By examining trends in the estimated redox potentials, we showed that NAD(P) has a reduction potential range that represents a near optimal adaptation given biochemical constraints and selection pressures imposed throughout evolution.

We find that NAD(P) can reversibly accommodate the reduction of a wide range of carbon functional groups, including activated carboxylic acids, and carbonyls, and can also support the common irreversible redox transformations of extended central metabolism–i.e., reduction of hydroxycarbons and oxidation of carbonyls to un-activated carboxylic acids.

We also find that the main cellular electron carrier, NAD(P), is 'tuned' to reduce the concentration of reactive (and potentially damaging) carbonyl functional groups, thereby keeping the cellular environment more chemically stable. Importantly, these insights were obtained thanks to the enhanced resolution provided by the large quantum chemistry dataset, which uncovered important patterns not accessible using traditional analyses.

Example 2: – In this work we analyzed the structure and thermodynamics of portions of "carbon redox chemical space": the chemical space of all possible redox states of linear-chain n-carbon compounds (for n=2-5) [63].

For every molecule in n-carbon redox chemical space, each carbon atom can be in one of four different oxidation levels: carboxylic acid, carbonyl (ketone or aldehyde), hydroxycarbon (alcohol), or hydrocarbon. The edges that connect molecules in redox chemical space represent reductions or oxidations that change the oxidation level of a single carbon atom.

After generating these redox chemical spaces, we found an interesting pattern: regardless of the number of carbon atoms considered, in every such redox chemical space there is always one and only one molecule with a maximal number of reactions (edges) connecting it to its oxidized or reduced products (i.e. maximal redox degree). This molecule is always the n-carbon aldose sugar (e.g. erythrose for n=4, ribose for n=5, glucose for n=6, etc.).

We then understood that this result can be understood through a simple and generalizable argument. The n-carbon aldose sugar satisfies the two constraints required to have the maximal number of redox connections:

- each atom must be in an "intermediate" oxidation level that can be both oxidized and reduced. Therefore all inner carbon atoms (i.e. atoms 2 and 3 in 4-carbon linear-chain molecules) must be in the hydroxycarbon oxidation level, while carbon atoms at the edges (i.e. atoms 1 and 4) can be either in the carbonyl (aldehyde) or hydroxycarbon oxidation level.

- The molecule must not be symmetric under a 180 degree rotation along its center. Thus the two edge atoms must be in different oxidation levels.

This leads uniquely to the aldose sugar molecular redox state configuration.

XVII. INSIGHT BY ERIN R. JOHNSON

I can provide two examples for you.

The first is from my graduate studies, when I was working toward development of our exchange-hole dipole moment (XDM) dispersion method. In an early iteration of the method, we only included C6 dispersion terms. When testing it on a benchmark set of molecular-dimer binding energies, we noted that, while the method worked well for most systems, the pi-stacked benzene dimers were distinct outliers [64].

After some investigation, this suggested to Axel Becke and I that we were missing some important component of dispersion physics, specifically inclusion of the higher-order C8 dispersion term arising from dipole-quadrupole interactions. Inclusion of both C8 and C10 in XDM improved performance for pi-stacked systems, giving a much more balanced treatment across the full set of intermolecular complexes [65].

We have since shown that inclusion of higher-order dispersion terms, particularly C8, is essential for accurate modeling of molecular crystals and of layered materials, such as graphite. This is summarized in my recent perspective [66].

A second example comes from a collaboration with Prof. Graeme Day at the University of Southampton. Graeme's group had performed DFT calculations on a set of 350 organic co-crystals to assess their stabilities [67].

During a seminar visit to Dalhousie, Graeme mentioned that 6 of these compounds were outliers, where DFT optimization lead to organic salts, rather than the expected co-crystals. This suggested to me that the density-functional delocalization error might be responsible and lead to a subsequent publication [68], which provided a dramatic demonstration that delocalization error can affect not only energies, but also structures, of molecular crystals.

There are possibly more examples from my research, but these are the two that immediately spring to mind.

XVIII. INSIGHT BY LYNN KAMERLIN

I see myself primarily as a biochemist/physical organic chemist, for me the simulations are a hypothesis testing tool, rather than a tool I expect to give me the answer. So I first sit down and think a lot about the problem, how to define it, what simulations would you need to test different (bio)chemical scenarios, and how can you address the problem in the best way. I also do a lot of validation, because most simulation approaches in my field have shortcomings, and can sometimes give results that really defy everything we know about biochemistry, which is usually a problem with how the problem was defined/simulations were set up, it worries me a lot that a lot of computational biochemists have absolute trust in the simulations as truth, and don't think about the limitations based on how the problem is set up and shortcomings in the methods. If there is an outlier, that deviates from what I would expect based on my knowledge and intuition as a (bio)chemist this is first and foremost a cause of serious concern for me, and a reason to get the students and myself to go back to the experimental data and look at it very carefully and start a massive debug process. Sometimes the simulations do really reveal something that would not have been obvious from the calculations, but I work on the "extraordinary claims require extraordinary evidence" premise so it takes a lot to convince me that this is new (bio)chemistry rather than a problem with the simulation setup. If we see it frequently enough over multiple systems in different settings, then I might believe it's a trend.

I actually have a nice example of this, although the revised draft will not be ready in time. We have been doing additional simulations and analysis on this preprint to quantify the ground-state destabilization we claim is really important: [69]

One of the things that was very confusing is that all our variants / compounds cleave preferentially the pro-(R) carboxylate group, which is what you would experimentally expect, except CLG-IPL which suddenly cleaves pro-(S). Lots of head scratching but then we saw it again and again with every compound. The preprint version has a lot of explanations for why the simulations may not be adequate, 6 mutations at once, no crystal structure etc. In the interim we have done a ton more simulations, looking at other things than just what we are looking at in the first round of the preprint. We have actually found out something cool they did during the experimental evolution, that they are not aware of, that caused this flip, and that is something that can actually be used for rational engineering (confidential right now, making final revisions to the revised manuscript). I think that's pretty neat in terms of a situation where everything points to the simulations being wrong or problematic, but actually it was something they had totally missed from their experimental setup and actually the outlier from the simulations was real, and not just real but something that can be manipulated on purpose for rational design once you know this is what the system is doing.

XIX. INSIGHT BY HEATHER J. KULIK

This is the closest anecdote I could think of from our own work:

Early on, we developed graph-based representations intended to systematically incorporate atom-by-atom contributions to property predictions in transition metal chemistry [70]. These feature sets were quite large and high-dimensional relative to our data set sizes. We were at first just interested to understand if we carried out feature selection if that improved model performance. We also wanted to identify best practices if we were interested in selecting features for multiple properties. Although it may be simple and intuitive in retrospect, what we observed strongly influenced the direction and focus of our research in accelerating chemical discovery in inorganic chemistry from that point on. Specifically, we noticed that interpreting selected features, while not truly unique, could help us to rationalize when our models performed well. For instance, the features selected to predict the spin of a molecule focused heavily on ligand-field-dependent aspects of the direct coordination environment, where ligand field theory is a well known phenomenological model to predict spin states that our feature selection had essentially revealed to us. But what was more interesting were cases where we had extremely limited prior experience and expectation about what features should be selected. We found that things like redox potential depended on completely different length-scales and properties (i.e., more size and coordination based than electronic/ligand field based), giving us a map and a path forward to carrying out orthogonal design of those two properties (expanded upon in [71]). Essentially, the maps of features allowed us to map and predict when properties should change in transition metal chemical space without requiring us to run the ML model we had trained or carry out the computation. Since then, we have used this approach to develop structure-property relationships and understanding across a range of challenges in energy storage (e.g., [72]) and catalysis (e.g., [73]), and the same design principles generally hold - for instance, redox-related properties are often used to rank expected performance of catalysts, but we know that when the catalytic property of interest is highly metal-directed,

that our feature maps will tell us that the relationship should not hold past the typical set of catalysts people traditionally include in a smaller scale study. We've used these guiding principles and representations to go back and interpret broader properties (e.g., gas separations in metal-organic frameworks) and shown how these features can be sensitive reporters of when researchers' lack of data set diversity can limit their conclusions [74]. Overall, we consistently find the extent to which selected features vary tells us how independently these properties can be tuned, giving us a path for when multi-objective design is most likely to be fruitful. Excitingly and very recently, we've also been able to show (work in preparation) that these selected feature maps are largely invariant to the electronic structure method chosen, suggesting they also have the promise of introducing an error cancellation that gives us robust design principles, even when the underlying physics-based model (for us, DFT) may not be expected to perform well.

For us, at least, Figure 11 in [70] was the jumping off point for everything I described.

XX. INSIGHT BY JEAN-PAUL MALRIEU

Tiny discoveries – Quantum Chemistry is a rather modest and applied discipline, at the border between physics and chemistry. It applies the basic equations of Quantum Physics, without questioning them (no epistemological revolution is to be expected from that practice), and since the objects it considers are molecules, in their infinite diversity, it concentrates on rather specific architectures and their properties. Chemistry is a science of differences and perturbations, often small ones; reaching new laws is difficult in this field.

"Discovery" is an intimidating word. Continents have been discovered, structures and laws too. And sometimes quite general and novel phenomena emerge, which deserve to be called discoveries, such as superconductivity. Aromaticity was proposed before its theoretical justification, and so it happened for Lewis electron pairing.

Looking at the story of quantum chemistry, I essentially see the Woodward and Hoffmann rules as a "discovery". Their work simultaneously builds concepts regarding the stereo-specificity of a wide class of reactions, formulates a law and provides an interpretation. The very interesting story of this discovery has been told in great detail by Seeman (J. Org. Chem., 80, 11632 (2015)), and it is clear that this work consisted first of the identification of a problem and then its rationalization. The computations (at this time very primitive) did not play a central role. Quantum Chemistry fills three roles:

- It solves specific problems, provides more and more precise values of given properties of given molecules or super-molecular architectures, it answers quantitative demands raised by experimentalists. It also becomes a tool for imagination of new, as yet unsynthetized frames possessing original properties, a tool for exploring the possible. But if we show the stability of a yet non-existing molecule, suggesting new horizons of the vast world of chemistry, might we speak of discoveries? Maybe, but we rarely are fortunate enough to play this game, and again, the forces at work would be imagination and logic, before a confirmation of our intuition by heavy computations.

- Quantum Chemistry provides accurate approximations to approach the solutions of Schrödinger's equations – this is the task of methodologists, a very satisfying activity (to which I dedicated most of my efforts). But the formulation of fundamental structures of the quantum Many-Body problem, the Linked Cluster theorem for instance, is due to physicists and is essentially logical. The elegant Coupled Cluster formalism comes from nuclear physics, and again is a logical proposal. Then come technicalities to transcribe these formal tools into efficient codes, a domain which our discipline definitely performs with impressive success.

- Quantum Chemistry provides interpretations, explanations through concepts and models, identification of effects and frames of understanding how these various effects combine, often competitively, sometimes cooperatively. This is our major task, indeed. It becomes interesting when the effects involved and the way they combine have some generality.

I perhaps may give an example of interpretation from my own practice, namely the rationalization (and prediction) of the structure of ionized rare-gas clusters. Ar2+ is spectroscopically well-studied, the delocalization of the hole between the two atoms explaining the bonding phenomenon. What about heavier clusters, the existence of which was attested by mass spectrometry? What are their structures? Does the hole delocalization, which is directional, explain their existence? We performed ab-initio computations, came up with reliable results, Ar3+ is linear, with a delocalization of the hole, consistent with a mono-electronic picture. In principle, the delocalization of the hole could go on along the same axis, but the most stable Ar4+ is not linear, it only exhibits a plane of symmetry; it looks like a flag, and the larger clusters exhibited at first glance strange geometries. Everything could be rationalized: the polarization of a neutral atom by the positive charge of Ar3+ is energetically more favorable than delocalizing the charge over 4 atoms. Polarization prevails over delocalization of the charge. And in Ar5+, two neutral atoms stand

at the same distance from the linear Ar3+, but are kept together, of course by dispersion forces. Adding other atoms one builds a crown of 5 atoms centered between two of the atoms of the Ar3+ stick, around the axis. Then a second crown is constructed, centered on the midpoint of the line joining the other couple of charged atoms, according to the hierarchy of effects: delocalization¿ polarization¿dispersion.

This a case where computation came first, delivering a priori bizarre geometries, provoking thinking. The pleasure, and the possibility to send a message, to tell a story, is in the second step, in the emergence of a simple and physically grounded rationalization. The understanding there has some generality, it is valid for all rare gases, but it remains confined to a very tiny land of molecular science. It would be indecent to call that a discovery! And notice that the theoretical knowledge of the physical effects at work in the problem was a prerequisite for understanding, the logic of constructing a hierarchy of physical effects could not emerge from only numbers provided by a sophisticated codes. Human intelligence, i.e. curiosity plus knowledge, remain crucial, even in these days, where AI is (appropriately) praised .

References: [75, 76]

XXI. INSIGHT BY ANAT MILO

We were working on predicting aldehyde deuteration ratios using different NHC organocatalysts. Initially, the features used to identify linear regression models were taken from the pre-activation catalysts' ground state structures, but we were getting overfitted models with multiple parameters and cross terms. The reaction pathway is presumed to pass through a Breslow intermediate, where the aldehyde is connected to the catalyst. So we extracted features from the intermediate and since it had two possible conformations we added to the dataset features from both the E and the Z conformer. We ended up finding a predictive model with one parameter from the catalyst structure and another from the E conformer. Interestingly, features from the Z conformer didn't provide a good model. Later and in the context of another project, we found that the reaction profile with the Z configuration had high-lying intermediates and transition states compared to the E. So here, the linear regression model pointed to the correct conformation without the need for an elaborate computation of the full reaction mechanism.

On a more anecdotal note regarding the same model, we were looking at a small dataset of only 14 catalysts (mind you, we have to prepare and test all the catalysts experimentally, so not entirely small from that perspective). Our aim was not to provide a predictive extrapolative model, but rather identify the mechanistic features that make for an effective catalyst. So we deemed it unwise to have a training and test set and just performed a 3-fold cross validation (repeated randomly over 500 trials). Then reality presented us with an opportunity to test how predictive this model was. The model gave a goodness-of-fit R2 value of around 0.8 with one catalyst that was a clear outlier without which the model gave an R2 value of around 0.95. So I asked the postdoc who ran the catalytic experiments whether there was something odd about that particular reaction. He said something in the lines of "now that you mention it, that reaction formed a precipitate once it was set up, which was not the case for other reactions". We asked the postdoc who had prepared all of the catalysts when this particular catalyst was synthesized and it turned out that it was a two year old batch that had somewhat decomposed. Once the catalyst was prepared again the resulting deuteration ratio actually fit the linear regression plot and we got a 0.94 R2 value. So even though this was not our goal, we ended up demonstrating to ourselves the power of mathematical models in identifying problematic reactions, but also, that our specific model was rather predictive.

The paper was recently uploaded as a preprint and is now under review: [77]

XXII. INSIGHT BY FRANK NOE

I am not sure if I have the kind of answers you are looking for, we are mostly developing methods. But here's two that I am particularly proud of and that also rely on generalizing ideas that have popped up by anecdotal observations:

1. Boltzmann Generators [78] – We work a lot on the sampling problem in many-body statistics. I had this idea that if we could find a clever variable transformation of state space, perhaps we could sample much more efficiently. That led me to invertible neural networks, where complex variable transformations could be mathematically represented and machine-learned. I got down and formulated this problem for statistical mechanics, where we often have an energy function or path action given with respect to which we want to sample, and it turns out this is in principle possible by just generating from the neural network and evaluating the energy-function pointwise. So the idea of the Boltzmann Generator was born and 6 months later we had a completely new approach to sample many-body systems in a principled way without relying on making tiny steps in state space.

2. PauliNet [79] – We had this vague idea of solving the electronic Schrödinger equation with a deep learning representation of wavefunctions. It became clear early that this would lead to a Quantum Monte Carlo method, but it wasn't clear how to best represent fermionic wavefunctions, and whether we could even reach a high enough accuracy to justify the com-

putational cost. The first approach to use deep learning for the Jastrow function worked in principle, but seemed just as an expensive way of what diffusion Monte Carlo already did. The key idea came when we found the backflow method - an approach introduced in the QMC literature, to transform orbitals to many-electron functions such that the Slater determinants built from them become more expressive. But backflow hasn't had much success in quantum chemistry so far. It struck us that this limited success what just because the functions used were not expressive enough, but this would be an ideal approach for optimizing them with Machine Learning. The idea for PauliNet was born. Now we can solve system really tricky quantum chemistry systems with 30 electrons to extremely high accuracy, at a cost of order (N^4). We hope to get in the regime of 100-200 electrons with this black-box method, and at that point there would be nothing on the market that can compete.

XXIII. INSIGHT BY JENS K. NØRSKOV

Discovery of scaling relations as a basis for a theory of transition metal heterogeneous catalysis –

Studying trends in adsorption energies and transition state energies for surface reactions made us realize that energy differences from one metal surface to the next scale with each other[80, 81]. The adsorption energy of OH scales with the adsorption energy of O and the adsorption energy of CH2CH3 scales with the adsorption energy of C (or CH2), for instance. Similarly, the transition state energy for N2 dissociation scales with the N adsorption energy and that of CO with the C and O adsorption energy. In fact, the interaction of a given molecule with a transition metal surface scales with the adsorption energy of whichever atoms form bonds to the surface. A subset of such correlations, known as Brønsted-Evans-Polayni (BEP) relations have been observed in many branches of chemistry. They typically relate an activation energy to a reaction energy (or, equivalently, the rate to the equilibrium constant) of an elementary reaction step. Yet, the generality was not discovered until we had reliable computational results that could span a large enough range of energies to reveal the trends.

It typically takes many (hundreds or thousands) of different intermediate and transition state energies to define the kinetics of a full catalytic reaction. This has made it extremely difficult to define some simple rules in this field, let alone design new catalysts on a scientific basis. The discovery of scaling relations has changed this drastically by projecting this extremely-multi-dimensional problem onto a few-dimensional one, allowing the kinetics of a full catalytic reaction to be expressed as a function of a few bond energies or descriptors in what is known as volcano plots. This has provided rationalization of a considerable part of heterogeneous catalysis[82]. It has also defined a set of catalyst design rules[83, 84].

XXIV. INSIGHT BY ARTEM R. OGANOV

Here is what fits closest:

Story 1 – Physicists know well that as you compress solids, they eventually transform into free-electron metals. Sodium, however, is already a nearly free-electron metal at normal conditions, and one might expect it to get only closer to the free-electron regime upon compression. We, however, predicted that at pressures of less than 2 million atmospheres sodium becomes... a transparent semiconductor. This was also confirmed by experiment, and a joint theoretical-experimental paper was published [85]. *(Note from authors: Ref. [85] is also discussed by Chris Pickard)*

We found this transparent sodium to be an electride. Many more pressure-induced electrides followed. And band gap opening under pressure was found in lithium (and its new phases are also electrides). The explanation of these phenomena is based on the important role of core electrons under pressure.

Story 2 – Solid metallic hydrogen, recently synthesized at the very high pressure of 4.25 million atmospheres, was long believed to be a room-temperature superconductor. It was hypothesized by Ashcroft, and later confirmed by theoretical and experimental works, that one can obtain high-Tc superconductors at lower pressures by alloying/reacting hydrogen with small amounts of other element(s). Indeed, the current record of superconductivity is LaH10, which has Tc = 250 K at the pressure of 1.7 million atmospheres. Many viewed this as electron-doped metallic hydrogen. However, we have found[86] that superconductivity of such hydrogen-rich hydrides very strongly depends on the electronic structure of the metal atom - so much so that even within otherwise very similar lanthanoids one observes huge differences in the superconductivity of hydrides. For example, LaH10 and CeH9 are high-Tc superconductors, but PrH9 is a low-Tc superconductor[87], and NdH9 is not a superconductor at all[88]. This means that these polyhydrides cannot be considered as forms of metallic hydrogen and gives a principle for designing new high-Tc superconductors. We have used this design principle to predict a number of hydride superconductors with Tc ¿ 200 K.

XXV. INSIGHT BY JUAN PEREZ-MERCADER

My work in science has always been driven by the mathematical/logical consistency underlying the phe-

nomenology and (since the beginning of the 1980s) the algorithmic representation of the observations and their computer representation. *These have led me to be a "reductionist" with the soul of a "non-reductionist" by consistently relying on first principles, their mathematical expression and the ensuing tension between experiment and computer simulation. Using the latter as a means to complement actual experimentation.*

In a way, I look at "physical laws" as an idealized human representation of reality which is valid up to a certain degree of fine-graining (outside-in), when new phenomena that were integrated in the preceding coarse grained (inside-out) description. Physical laws as represented by mathematical equations are what guides my discovery. Then I use computer solutions to inform experiment and when I can, do the experiments to check the validity of the laws.

I have applied the above thinking in Grand Unified Theories to study proton decay. The current data is compatible with my predictions of 1980, but not enough data yet.

Also in supersymmetry/supergravity extensions of the Standard Model of Matter, where our mathematical calculations were used to pin down parameter values to hunt for the Higgs particle. The big data from the CERN experiments where analyzed in the light of what the extensions of the mathematical/theoretical Standard Model suggested. And ... bingo, they discovered the Higgs particle and bound their fundamental values.

In studying galaxies at large scales, I used another very important principle: universality (cf. Wilson, Gell-Mann and Kadanoff). By establshing analogies between analogies (cf. Stephan Banach) and using universality, together with lots of astrophysical data which constrained the galaxy-to-galaxy correlation to a power law, we were able to predict the value of the exponent.

In the study of the Lense-Thirring effect, we were looking for an "outlier" in the computer simulations of the Earth-LAGEOS satellites System and/or in the data. If there was an outlier it would represent a departure from Einstein General Theory of Relativity. We not only find any outliers, we found that Einstein's theory was "Queen" and that its prediction of the Lense-Thirring parameters was perfect. This has been vindicated in recent orservations/computer simulations of binary stars and pulsars.

Relying on ideas from universality and first-principles, in the early 2000s I began to look for a compact and simple representation of the properties common to all living systems. Using computer generated solutions we saw that certain "simple" sets of stochastic reaction-diffusion equations represented them. The amounts of data generated were very large (we had to build a beowulf cluster to do this) and the analysis required insightful (i.e., guided by their representation of underlying principles, such as "handling of information") studies of the parameters. There were segregated regions of parameter space which contained the interesting phenomenologies. The study of the numerical data and its correspondence with first principles, led me to conceive (a) that there would exist a representation of life which did not use biochemistry if (b) non-biochemical chemistry computed. The form of the equations led me, after a considerable effort searching in databases for standard chemical kinetics, that the Belousov-Zhabotinsky (BZ) reaction played a fundamental role in these mimics. Therefore I began to think (around 2005-6) that one should build an advanced chemical automaton using BZ. Based on this "knowledge" we built a chemical Turing machine (patent 2017, papers in 2019) and are now able to build in-vitro totally non-biochemical life mimics, and are also on our way to build an "Avogadrian" computer.

To summarize, in my research I combine first principles, logical consistency (as represented by mathematical expression of the first first principles) and computer simulations to inform experiments which are then fed back to the theoretical description. So, if I wanted to build a plane I would not imitate a bird or a flying insect. Instead I would understand the principles involved in flying and implement them ex-novo. Just like the Wright brothers did!

XXVI. INSIGHT BY CHRIS PICKARD

This is a great question, and one close to my heart.

Ab Initio Random Structure Searching (AIRSS) [89, 90] involves the high throughput first principles relaxation of diverse stochastically generated structures (crystals, clusters, surfaces, interfaces). The emphasis is very much on 'exploration', and hunting for these outliers, or surprises. I try to highlight the new phenomena uncovered in the searches, rather than the details of the crystal structure. But, of course, you have to get those right to make meaningful predictions. When I find a surprising result, I put a lot of effort in trying to get it to go away (most do). I think this has lead to a very high success rate, where the predictions have found to be good.

Many of the early applications were to the high pressure sciences, starting out looking for superconductivity and metallicity in the hydrides.[89] This has grown to be a very active area with well known successes.[91]

I think the best examples of what you are looking for are the following:

Mixed phases in hydrogen [92] – In the course of trying to understand Phase III of dense hydrogen (our prediction of C2c-24 appears to be standing the test of time [93]) I was confronted by a stunning (I had to go and sit on the beach in St Andrews for a while

to recover - it was a good summer) metastable structure of a type not previously suggested for an element. It consisted of layers, alternating between graphene-like and molecular. I felt these structures must be important for dynamically stabilised phases (ZPE, or T), but the techniques were not then ready to permit a full phase diagram to be computed. Nevertheless, we published the structures as part of our study, and emphasised them in presentations to experimentalists. Interestingly, at the time it was as if they couldn't see them - they didn't answer any questions they had and "muddied an already murky field". This changed when Goncharov and Gregoryanz contacted me with a puzzle - they were seeing a surprising softening in a Raman peak in warm (room temperature) hydrogen. I suggested that they were likely seeing these mixed phases, and this indeed was the case.[94] The mixed phases are now an established feature of the hydrogen phase diagram. I think it is fair to say that, given the experimental challenges in determining the positions of protons, our current understanding of dense hydrogen is to a large extent due to first principles structure searches, with much having been mapped out in Ref. [92].

A question could be asked - why were we so successful. Of course, high throughput searches made a big difference, but these structures could probably have been found using MD. I think the reason that they were not is because MD is often conducted in cubic unit cells, and fixed numbers of atoms. Typically multiples of 8. But my candidates for dense hydrogen all required multiples of 12. I had been in the habit of not assuming the number of atoms in the unit cell, and choosing them randomly as part of the structure generation. This was also very important in the aluminium case below, and highlights the importance of unbiased random searches.

Ionic ammonia [95] – In searching for molecular crystal structures a well established protocol is to stochastically pack the molecular units. This will shrink the search space, and dramatically increase the odds of finding low energy configurations. But this might be at the cost of missing the most stable! Following my habit of assuming as little as is computationally feasibly, I had been searching for dense phases of NH3 by individually throwing the N and H atoms into randomly shaped unit cells. It was a fairly routine project, but I was jolted awake one early morning on checking the most recent results. I was convinced something was broken - the most stable units (by some margin) where NH2 and NH4, not NH3. This possibility has not been discussed previously, and it was not something we were looking for. After careful testing, the result held, and the spontaneous ionisation of NH3 is now an established experimental fact,[96] and spontaneous ionisation more generally is considered as a possibility where it might not have been previously.

Complex phases of aluminium at terapascal pressures [97] – We (and others, in particular Yanming Ma) had starting to find a great number of electride type structures in the dense elements. [85, 98] *(Note from authors: Ref. [85] is also discussed by Artem R. Oganov)* One striking feature of these were the localisation of states under increasing pressure, and band narrowing. I wondered whether I could find a non-magnetic element that under the right conditions would exhibit magnetism. So, I begin the hunt, systematically working my way through the periodic table. Importantly, it turned out, I was randomly choosing the number of atoms in the unit cell. When it came to aluminium I was shocked (again) to find the most stable structure at 3 TPa (it was trying a wide range of pressures, and my expertise in generating pseudopotentials meant I could reach pressures those those dependent on supplied potentials could not) contained 11 atoms in the unit cell. Few groups would even consider odd numbers of atoms as a possibility. It was a lot more stable that the other candidates, and initially, when I visualised it, it made no sense. It looked to be amorphous, or still random somehow. But I kept spinning the structure around, making supercells, in the visualiser, and eventually all became clear. The structure consisted of tubes and chains of atoms. I was well aware of the work of Nelmes and McMahon on host guest phases in the alkali metals - Volker Heine had publicised it in Cambridge. And this was exactly what I was seeing - an approximant of a kind of 1D quasicrystal. Once I had seen that, it was straightforward to construct other, larger, approximants, and estimate the ideal lattice parameters for the host and guest phases. I was also able to determine that the structure was of the electride type, and construct a simple model for it.

This result has not been confirmed experimentally - yet. But it has had a large impact on the field - it showed that materials under extreme compression might be complex, and not just simply close packed in some way. This has inspired the high pressure community, in particular the shock physicists, for example being used as part of the justification for using the NIF to perform exploratory science. Continuing my sweep through the periodic table, I did eventually manage to find magnetism in an electride phase, in potassium.[99]

XXVII. INSIGHT BY MARKUS REIHER

1) In this work [100] we combined our fast interactive quantum mechanics engine (reported here [101]) with the VR framework Narupa of Dave Glowacki and his group which enabled interactive exploration of chemical reaction space (by virtue of high-throughput computation in real time) up to the point where we could use the data flow to train a neural network that learns the potential energy surface. This new setting

that combines ultrafast quantum methods for high-throughput calculations with virtual reality and machine learning defines a conceptually new approach toward the exploration of the molecular world. Interestingly, the result for human perception of the (quantum) molecular world is not easy to describe in sentences due to a lack of standard concepts to report such immersive experiences. Instead, it requires one to try it out in order to understand its potential.

2) Our work on automated screening of reaction mechanisms demonstrated that a full predictive understanding of chemical reactions can only be achieved if all relevant species and connecting transition states are explicitly evaluated. This high-throughput quantum approach turns the study of basically all chemical reactions into a big data problem. The wealth of data, automatically processed by evaluation algorithms has allowed us to show that there exist numerous side pathways in a catalytic process that need to be taken into account. This general insight was gained for a typical homogeneous catalyst, the nitrogen fixation catalyst by Yandoulov and Schrock reported here[102].

Or in another example (the formose reaction) we understood that an exploration in full conformational depth is necessary to uncover all relevant reaction pathways,[103] Those do not necessarily need to start from the lowest-energy conformers as other conformers, higher in energy, may actually feature lower barriers, an insight of far-reaching consequences; see Fig. 4 of that paper. Interestingly, this problem appears to be not of the kind that one can extract a new conceptual general idea from, but all our calculations point to an understanding of a necessity of explicit calculation of the full many-particle wealth of chemistry in order to figure out what is going on in a quantitative sense. Of course, this is intimitaley connected to the peculiarities of chemical reaction mechanisms and not related to physical properties in general.

XXVIII. INSIGHT BY JEAN-LOUIS REYMOND

Thanks for your mail, this is an interesting inquiry! To your questions:

on our side our big data project has been to understand chemical space by enumerating all possible molecules from first principles. This generated billions of molecules, and we quickly came to the limits of what computer allow to do. The next question was to do something useful with these billions of molecules - we had to look into them in a clever way, and we turned to mapping chemical space to understand the content of chemical space and search for bioactive compounds. Like everyone else, we were a bit scared of big numbers - combinatorial chemistry is dominated by the fear of missing something. What's come out of visualizing chemical space was unexpected: we can actually get an overview and proceed with insights. In fact, we use the computer not only to generate the data, but also to help us see the data, because we cannot leave it to the machine to solve the problem entirely. Also I find that bench chemists only commit to a difficult project if they can decide, so they need to choose, not the computer. At present we are implementing AI in that game at all levels: enumeration, visualization, activity prediciton and retrosynthesis. Whether this will change everything is open - it's exciting to ask because things a possible with AI that were simply not even imaginable previously.

we've done some small molecule drug discovery with our GDB, there is some ongoing unpublished, a previous study with a lot of details [104]. On the side of application, we have used our chemical space approach recently to discover antimicrobial peptides [105]

XXIX. INSIGHT BY STEFANO SANVITO

Let me give you two examples from my group. I am not sure they will match in full the brief ... I'll let you decide.

1) Discovery of new magnets – We have carried out a massive high-throughput search (DFT - PBE) for novel magnets, crystallising in the Heusler alloys structure (about 500,000 prototypes calculated). For all of them we have assessed the thermodynamical stability (convex hull) and their basic magnetic properties. We have evaluated their possible Curie temperature via a machine-learning model trained on experimental data. With this we have identified a new (never made before) magnet with extremely high T_C (predicted 938K), Co2MnTi. This has then be made in the lab (successfully) and the T_C was measured at 940 K. In my opinion it is a kind of big deal since: 1) it is one of the few example of computer-to-lab pipeline, 2) only 5% of the magnets have T_C larger then 600 K, 3) it is the only example I know of the computer "design" of a material with macroscopic magnetic order (magnetism). All this has been published here: [106]

2) Designing rule for magnetic anisotropy in molecular magnet – We have constructed a SNAP force field (similar to GAP, but with linear regression) for Co-based single molecule magnets. With the same structural descriptor we have constructed a "force field" for the magnetic anisotropy tensor. Then we have defined a "functional" containing the energy and the magnetic anisotropy. This has been minimised over the entire configuration space (30 atoms - 90 degrees of freedom) in order to find the main structural parameters of the molecule driving the magnetic anisotropy. In particular we find that the alteration of a bond angle and one bond length determine the anisotropy and changes by less than 5% produces a change in anisotropy of more than 50%. All this is

published here: [107]

XXX. INSIGHT BY FRANZISKA SCHOENEBECK

we have several such examples where a computational discovery/unusual result led us to dig deeper and eventually we proved the new phenomenon experimentally. The following three especially stand out in this context:[108–110]

XXXI. INSIGHT BY ILJA SIEPMANN

Here is another computation-led work that via validation led to a patent[111]. We also tried to get a patent for the hydrocarbon isomerization, but it turned out that Chevron had patents for these materials but had given them different names.

XXXII. INSIGHT BY ALEX SODT

This is from our paper "The molecular structure of the liquid ordered phase of lipid bilayers" [112]

Experimentally, three-component mixtures of cholesterol, an unsaturated lipid, and a saturated lipid form phase-separated two dimensional liquids. We used the "Anton" special purpose molecular dynamics computer to run a simulation of a mixture predicted to phase separate. We applied hidden Markov modeling to assign the state to one of the two co-existing phases (this was the hidden state). Our HMM observable was the local lipid composition (one phase is highly enriched in the saturated lipid and so lipids in this phase will be near other saturated lipids). This allowed us to compare to NMR order parameters (critically, without using the order parameter information to assign phase).

The assignment let us quantitatively analyze the properties of the phase under co-existence conditions. We verified the structure using a simulation of the pure liquid ordered phase (not coexisting conditions). The method has been implemented by a number of other groups looking for similar information.

It's clear that this doesn't strictly qualify under the request you made since the main consequence of the tool was the assignment of lipid phase (ordered or disordered) and the most important thing we learned could have been learned from the simulation of the pure ordered phase without the HMM. However, it was critical to observe the same structure under co-existing liquid conditions.

XXXIII. INSIGHT BY ISAAC TAMBLYN

To be honest, I'm not sure I've actually done what you described in any AI / ML paper so far. We've certainly developed tools which are faster, and I've seen some AI produce results which were not obvious before hand, but I'm not sure I got a general understanding from it.

The closest example I can think of is in this paper, where the neural network learned how to produce to different structures in self assembly [113] I would have had no idea how to even attempt this before hand.

XXXIV. INSIGHT BY DONALD TRUHLAR

An important problem in electron transfer and in fact in photochemistry in general is diabatization, i.e., calculating diabatic states. We have worked on this for many years, and the best methods are orbital-based and based on configurational uniformity. One needs localized orbitals or diabatic molecular orbitals, and one needs to identify diabatic prototype state functions.

This year, we showed that we can get diabatic states by a deep neural network. We do not need to find diabatic molecular orbitals; we do not need to identify diabatic prototype configuration state functions. The deep neural network finds the diabatic states with minimal help from humans.

We have one completed paper on this [114]. We also have more good results that will be written up in a second paper and we are starting to apply it to harder problems. We are enthusiastic about the new method.

XXXV. INSIGHT BY ALEXANDRE TKATCHENKO

Thanks for asking. I have a counterexample to your question :). This concerns the usage of data analysis / ML for breaking established textbook "laws". Namely, in chemistry there are many established quasi-linear relations. For example, molecular polarizability is supposed to be inversely proportional to the HOMO-LUMO gap among many other examples. When you generate enough data in chemical space with the purpose of constructing ML models you realize that all those "textbook rules" do not hold in large enough chemical spaces.

This paper provides an initial indication of such lack of simple "chemical correlations", [115].

Our recent review provides a more detailed analysis of this situation [116].

XXXVI. INSIGHT BY KOJI TSUDA

Molecular design is best done when humans collaborate with AI. But finding a narrative for such a paper is hard, because it is difficult to show that AI was necessary in developing the molecule. There are several on-going projects that humans select good molecules from a large number of ChemTS-generated molecules, and try to generate their modified versions. We still do not have published results.

We found that n-pi* excitation is used by AI to achieve desired absorption wavelength on our ACS central science paper [117], when humans mainly look at HOMO-LUMO excitation. It is an example that AI took a fundamentally different path, but it may not be something you wanted to hear.

XXXVII. INSIGHT BY ALEXANDRE VARNEK

Several years ago, we've published a paper [118] in which the outliers analysis was used to assess a quality of compounds studied experimentally by our partners. In particular, we've discovered that almost all detected ouliers correspond to hydrolysed or degraded compounds. This didn't lead to discovery of a new physical phenomena which you were looking for but, up to me, useful practical observation.

XXXVIII. INSIGHT BY TEJS VEGGE

1) Using an in house high-throughput genetic algorithm-DFT approach we analyzed ¿100.000 mixed metal halides for ammonia storage. This study outlined two new types ternary metal halide outliers which followed a different ab/desorption profile [119]. This insight let to a simple design rule that was used to design new and improved ammonia storage materials for industrial applications.

2) We performed a large DFT-level screening study of doped 2D and 3D layered oxyhydroxide catalysts for the oxygen evolution reduction reaction, finding that some of these dopants could induce a shift in the mechanism for the rate limiting step and lead to a reduction in the overpotential[120]. This insight was later applied by an experimental group to make and confirm that Zn-CoOOH followed a different mechanism with a lower overpotential than CoOOH [121].

XXXIX. INSIGHT BY ANATOLE VON LILIENFELD

1) Understanding of representations through experimentation – When looking at learning curves of energy ML models throughout compound space (test error vs training set size) we realized that we could raise the off-set by making our representation less *physical*. More specifically, using interatomic functions which grow linearly with distance (rather than decrease as energetic interaction do) the off-set of the learning curve increases. Using functions which grow quadratically, the off-set increases even more! From that we concluded that using a representation based on functional forms similar to interatomic force fields should yield the most data-efficient ML models. This "discovery" was published in "Understanding molecular representations in machine learning: The role of uniqueness and target similarity" [122] and the insight led to the subsequent development of the FCHL representation [123], currently one of the (if not the) most data-efficient representations for QML models applicable throughout chemical compound space.

2) Discoveries from outliers – We used ML models to predict formation energies of all the 2 M crystals one could generate using main group elements and the elpasolite crystal structure. We realized that there are many crystals with relatively low formation energy which would have non-zero lowest possible total oxidation numbers (when you add those oxidation numbers from all the possible ones tabulated in text-books which result in the smallest total value). This suggested that hitherto unknown oxidation numbers should exist! In particular, using this criterion, and based on a convex hull analysis, we could identify the NFAl2Ca6 crystal (out of the 2m) which was (a) expected to be stable and (b) contained an Al atom with a negative oxidation number (absent from all text book entries), which we validated using DFT (Voronoi, Bader, and Hirshfeld charge analysis). The insight from that is that there might be many more possible oxidation states which have not yet been considered. This was published in "Machine Learning Energies of 2 M Elpasolite (ABC2D6) Crystals" [124].

XL. INSIGHT BY EVA ZUREK

I will discuss the search for high-temperature superconductivity in hydrogen rich materials under pressure. I attach a couple of review articles I wrote on this topic [125–127].

After various groups had used crystal structure prediction to find the most stable structures of binary hydrides and calculate their properties it became pretty evident that the highest Tc binary materials would contain hydrogen and either an alkaline/rare earth metal, or a main group element. This inspired my group to carry out calculations on ScHn and PHn phases under pressure.

Regarding the alkaline/rare earths, it also became evident that various hydrogenic motifs (H-, H3-, H2, 1D chains, 3D clathrate cages) can be found in the lattices. Phases that contain H- or H3- become metal-

lic b/c of pressure-induced band overlap so they are not good metals and will not be good superconductors. Systems with quasi-molecular H2 units or other odd molecular motifs have a higher DOS at the Fermi level and are therefore better superconductors. But the highest Tc materials have "clathrate-like" hydrogenic cages. The computational predictions of high-Tc superconductivity in LaH10 (by Pickard/Ma and Hemley) were confirmed by two experimental groups. Bummer- my group chose to study ScHn instead of LaHn. Our predictions for Sc are yet to be experimentally confirmed.

We are currently trying to predict the structures of ternary hydrides that could be high Tc superconductors. By now I can look at a predicted structure and guess what its Tc is likely to be simply by looking at the structural motifs present in its hydrogenic lattice. The goal is to find a ternary that can be quenched to lower pressures (metastable) with an appreciable Tc.

Another system that was found to be a high temperature superconductor is H3S, which is made up of two interpenetrating sublattices. It has been realized that one can remove one of these lattices and stuff the cubes with other molecules. We have looked at methane (see attached manuscript on CSH7). The superconductivity is derived from the H3S lattice. We are looking for other molecules that could be intercalants.

[1] Borna Zandkarimi and Anastassia N Alexandrova, "Surface-supported cluster catalysis: Ensembles of metastable states run the show," Wiley Interdisciplinary Reviews: Computational Molecular Science **9**, e1420 (2019).

[2] Eric T Baxter, Mai-Anh Ha, Ashley C Cass, Anastassia N Alexandrova, and Scott L Anderson, "Ethylene dehydrogenation on pt4, 7, 8 clusters on al2o3: Strong cluster size dependence linked to preferred catalyst morphologies," ACS Catalysis **7**, 3322–3335 (2017).

[3] Avik Halder, Mai-Anh Ha, Huanchen Zhai, Bing Yang, Michael J Pellin, Sönke Seifert, Anastassia N Alexandrova, and Stefan Vajda, "Oxidative dehydrogenation of cyclohexane by cu vs pd clusters: selectivity control by specific cluster dynamics," ChemCatChem **12**, 1307–1315 (2020).

[4] Juan M Venegas, Zisheng Zhang, Theodore O Agbi, William P McDermott, Anastassia Alexandrova, and Ive Hermans, "Why boron nitride is such a selective catalyst for the oxidative dehydrogenation of propane," Angewandte Chemie International Edition **59**, 16527–16535 (2020).

[5] Mai-Anh Ha, Eric T Baxter, Ashley C Cass, Scott L Anderson, and Anastassia N Alexandrova, "Boron switch for selectivity of catalytic dehydrogenation on size-selected pt clusters on al2o3," Journal of the American Chemical Society **139**, 11568–11575 (2017).

[6] Timothy J Gorey, Borna Zandkarimi, Guangjing Li, Eric T Baxter, Anastassia N Alexandrova, and Scott L Anderson, "Coking-resistant sub-nano dehydrogenation catalysts: Pt n sn x/sio2 (n= 4, 7)," ACS Catalysis **10**, 4543–4558 (2020).

[7] Elisa Jimenez-Izal, Ji-Yuan Liu, and Anastassia N Alexandrova, "Germanium as key dopant to boost the catalytic performance of small platinum clusters for alkane dehydrogenation," Journal of Catalysis **374**, 93–100 (2019).

[8] Elisa Jimenez-Izal, Huanchen Zhai, Ji-Yuan Liu, and Anastassia N Alexandrova, "Nanoalloying mgo-deposited pt clusters with si to control the selectivity of alkane dehydrogenation," ACS Catalysis **8**, 8346–8356 (2018).

[9] Mai-Anh Ha, Jonny Dadras, and Anastassia Alexandrova, "Rutile-deposited pt–pd clusters: A hypothesis regarding the stability at 50/50 ratio," ACS Catalysis **4**, 3570–3580 (2014).

[10] Borna Zandkarimi, Timothy J Gorey, Guangjing Li, Julen Munarriz, Scott L Anderson, and Anastassia N Alexandrova, "Alloying with sn suppresses sintering of size-selected subnano pt clusters on sio2 with and without adsorbates," Chemistry of Materials **32**, 8595–8605 (2020).

[11] Borna Zandkarimi, Geng Sun, Avik Halder, Soenke Seifert, Stefan Vajda, Philippe Sautet, and Anastassia N Alexandrova, "Interpreting the operando xanes of surface-supported subnanometer clusters: When fluxionality, oxidation state, and size effect fight," The Journal of Physical Chemistry C **124**, 10057–10066 (2020).

[12] Borna Zandkarimi and Anastassia N Alexandrova, "Can fluxionality of subnanometer cluster catalysts solely cause non-arrhenius behavior in catalysis?" The Journal of Physical Chemistry C **124**, 19556–19562 (2020).

[13] Borna Zandkarimi and Anastassia N Alexandrova, "Dynamics of subnanometer pt clusters can break the scaling relationships in catalysis," The journal of physical chemistry letters **10**, 460–467 (2019).

[14] Zisheng Zhang, Borna Zandkarimi, and Anastassia N Alexandrova, "Ensembles of metastable states govern heterogeneous catalysis on dynamic interfaces," Accounts of chemical research **53**, 447–458 (2020).

[15] Lorenzo Casalino, Zied Gaieb, Jory A Goldsmith, Christy K Hjorth, Abigail C Dommer, Aoife M Harbison, Carl A Fogarty, Emilia P Barros, Bryn C Taylor, Jason S McLellan, et al., "Beyond shielding: the roles of glycans in the sars-cov-2 spike protein," ACS Central Science **6**, 1722–1734 (2020).

[16] Christopher D Wassman, Roberta Baronio, Özlem Demir, Brad D Wallentine, Chiung-Kuang Chen, Linda V Hall, Faezeh Salehi, Da-Wei Lin, Benjamin P Chung, G Wesley Hatfield, et al., "Computational identification of a transiently open l1/s3 pocket for reactivation of mutant p53," Nature communications **4**, 1–9 (2013).

[17] Benjamin P MacLeod, Fraser GL Parlane, Thomas D Morrissey, Florian Häse, Loïc M Roch, Kevan E Dettelbach, Raphaell Moreira, Lars PE Yunker, Michael B Rooney, Joseph R Deeth, et al., "Self-driving laboratory for accelerated discovery of thin-film materials," Science Advances **6**, eaaz8867 (2020).

[18] Tracy H Schloemer, Jeffrey A Christians, Joseph M Luther, and Alan Sellinger, "Doping strategies for small molecule organic hole-transport materials: impacts on perovskite solar cell performance and stability," Chemical Science **10**, 1904–1935 (2019).

[19] Ji-Youn Seo, Hui-Seon Kim, Seckin Akin, Marko Stojanovic, Elfriede Simon, Maximilian Fleischer, Anders Hagfeldt, Shaik M Zakeeruddin, and Michael Grätzel, "Novel p-dopant toward highly efficient and stable perovskite solar cells," Energy & Environmental Science **11**, 2985–2992 (2018).

[20] Qiong Wang, "Influence of a cobalt additive in spiroometad on charge recombination and carrier density in perovskite solar cells investigated using impedance spectroscopy," Physical Chemistry Chemical Physics **20**, 10114–10120 (2018).

[21] Nina Taherimakhsousi, Benjamin P MacLeod, Fraser GL Parlane, Thomas D Morrissey, Edward P Booker, Kevan E Dettelbach, and Curtis P Berlinguette, "Quantifying defects in thin films using machine vision," NPJ Computational Materials **6**, 1–6 (2020).

[22] Karl T Debiec, David S Cerutti, Lewis R Baker, Angela M Gronenborn, David A Case, and Lillian T Chong, "Further along the road less traveled: Amber ff15ipq, an original protein force field built on a self-consistent physical model," Journal of chemical theory and computation **12**, 3926–3947 (2016).

[23] Rebecca J Swett, Angela Elias, Jeffrey A Miller, Gregory E Dyson, and G Andrés Cisneros, "Hypothesis driven single nucleotide polymorphism search (hydn-snp-s)," DNA repair **12**, 733–740 (2013).

[24] Antonia A Nemec, Korie B Bush, Jamie B Towle-Weicksel, B Frazier Taylor, Vincent Schulz, Joanne B Weidhaas, David P Tuck, and Joann B Sweasy, "Estrogen drives cellular transformation and mutagenesis in cells expressing the breast cancer–associated r438w dna polymerase lambda protein," Molecular Cancer Research **14**, 1068–1077 (2016).

[25] Pavel Silvestrov, Sarah J Maier, Michelle Fang, and G Andrés Cisneros, "Dnarcdb: A database of cancer biomarkers in dna repair genes that includes variants related to multiple cancer phenotypes," DNA repair **70**, 10–17 (2018).

[26] Alice R Walker, Pavel Silvestrov, Tina A Müller, Robert H Podolsky, Gregory Dyson, Robert P Hausinger, and Gerardo Andrés Cisneros, "Alkbh7 variant related to prostate cancer exhibits altered substrate binding," PLoS computational biology **13**, e1005345 (2017).

[27] Nicole M Antczak, Alice R Walker, Hannah R Stern, Emmett M Leddin, Carl Palad, Timothy A Coulther, Rebecca J Swett, G Andrés Cisneros, and Penny J Beuning, "Characterization of nine cancer-associated variants in human dna polymerase κ," Chemical research in toxicology **31**, 697–711 (2018).

[28] Mark A Hix, Lai Wong, Ben Flath, Linda Chelico, and G Andrés Cisneros, "Induced mutagenesis by the dna cytosine deaminase apobec3h haplotype i protects against lung cancer," bioRxiv (2020).

[29] James TA Jones, Tom Hasell, Xiaofeng Wu, John Bacsa, Kim E Jelfs, Marc Schmidtmann, Samantha Y Chong, Dave J Adams, Abbie Trewin, Florian Schiffman, et al., "Modular and predictable assembly of porous organic molecular crystals," Nature **474**, 367–371 (2011).

[30] Angeles Pulido, Linjiang Chen, Tomasz Kaczorowski, Daniel Holden, Marc A Little, Samantha Y Chong, Benjamin J Slater, David P McMahon, Baltasar Bonillo, Chloe J Stackhouse, et al., "Functional materials discovery using energy–structure–function maps," Nature **543**, 657–664 (2017).

[31] Catherine M Aitchison, Christopher M Kane, David P McMahon, Peter R Spackman, Angeles Pulido, Xiaoyan Wang, Liam Wilbraham, Linjiang Chen, Rob Clowes, Martijn A Zwijnenburg, et al., "Photocatalytic proton reduction by a computationally identified, molecular hydrogen-bonded framework," Journal of Materials Chemistry A **8**, 7158–7170 (2020).

[32] Peng Cui, David P McMahon, Peter R Spackman, Ben M Alston, Marc A Little, Graeme M Day, and Andrew I Cooper, "Mining predicted crystal structure landscapes with high throughput crystallisation: old molecules, new insights," Chemical science **10**, 9988–9997 (2019).

[33] Siwar Chibani and François-Xavier Coudert, "Systematic exploration of the mechanical properties of 13621 inorganic compounds," Chemical science **10**, 8589–8599 (2019).

[34] François-Xavier Coudert, "Systematic investigation of the mechanical properties of pure silica zeolites: stiffness, anisotropy, and negative linear compressibility," Physical Chemistry Chemical Physics **15**, 16012–16018 (2013).

[35] Jack D Evans and François-Xavier Coudert, "Predicting the mechanical properties of zeolite frameworks by machine learning," Chemistry of Materials **29**, 7833–7839 (2017).

[36] Jarosław M Granda, Liva Donina, Vincenza Dragone, De-Liang Long, and Leroy Cronin, "Controlling an organic synthesis robot with machine learning to search for new reactivity," Nature **559**, 377–381 (2018).

[37] Jonathan Grizou, Laurie J Points, Abhishek Sharma, and Leroy Cronin, "A curious formulation robot enables the discovery of a novel protocell behavior," Science advances **6**, eaay4237 (2020).

[38] Juan Manuel Parrilla Gutierrez, Trevor Hinkley, James Ward Taylor, Kliment Yanev, and Leroy Cronin, "Evolution of oil droplets in a chemorobotic platform," Nature communications **5**, 1–8 (2014).

[39] Carl A Fogarty, Aoife M Harbison, Amy R Dugdale, and Elisa Fadda, "How and why plants and human n-glycans are different: Insight from molecular dynamics into the "glycoblocks" architecture of complex carbohydrates," Beilstein journal of organic chemistry **16**, 2046–2056 (2020).

[40] Aoife M Harbison, Lorna P Brosnan, Keith Fenlon, and Elisa Fadda, "Sequence-to-structure dependence of isolated igg fc complex biantennary n-glycans: a

molecular dynamics study," Glycobiology **29**, 94–103 (2019).

[41] Daniel Schwalbe-Koda, Zach Jensen, Elsa Olivetti, and Rafael Gómez-Bombarelli, "Graph similarity drives zeolite diffusionless transformations and intergrowth," Nature materials **18**, 1177–1181 (2019).

[42] Sebastian Mai, Marvin Pollum, Lara Martínez-Fernández, Nicholas Dunn, Philipp Marquetand, Inés Corral, Carlos E Crespo-Hernández, and Leticia González, "The origin of efficient triplet state population in sulfur-substituted nucleobases," Nature communications **7**, 1–8 (2016).

[43] Christian Reichardt, Cao Guo, and Carlos E Crespo-Hernández, "Excited-state dynamics in 6-thioguanosine from the femtosecond to microsecond time scale," The Journal of Physical Chemistry B **115**, 3263–3270 (2011).

[44] Christian Reichardt and Carlos E Crespo-Hernández, "Room-temperature phosphorescence of the dna monomer analogue 4-thiothymidine in aqueous solutions after uva excitation," The Journal of Physical Chemistry Letters **1**, 2239–2243 (2010).

[45] Marvin Pollum and Carlos E Crespo-Hernández, "Communication: The dark singlet state as a doorway state in the ultrafast and efficient intersystem crossing dynamics in 2-thiothymine and 2-thiouracil," The Journal of chemical physics **140**, 071101 (2014).

[46] Marvin Pollum, Steffen Jockusch, and Carlos E Crespo-Hernández, "Increase in the photoreactivity of uracil derivatives by doubling thionation," Physical Chemistry Chemical Physics **17**, 27851–27861 (2015).

[47] Marvin Pollum, Steffen Jockusch, and Carlos E Crespo-Hernández, "2, 4-dithiothymine as a potent uva chemotherapeutic agent," Journal of the American Chemical Society **136**, 17930–17933 (2014).

[48] Marvin Pollum, Lara Martinez-Fernandez, and Carlos E Crespo-Hernandez, "Photochemistry of nucleic acid bases and their thio-and aza-analogues in solution," Photoinduced phenomena in nucleic acids i , 245–327 (2014).

[49] Victoria Vendrell-Criado, Jose A Sáez, Virginie Lhiaubet-Vallet, M Consuelo Cuquerella, and Miguel A Miranda, "Photophysical properties of 5-substituted 2-thiopyrimidines," Photochemical & Photobiological Sciences **12**, 1460–1465 (2013).

[50] Lara Martínez-Fernández, Leticia González, and Inés Corral, "An ab initio mechanism for efficient population of triplet states in cytotoxic sulfur substituted dna bases: the case of 6-thioguanine," Chemical Communications **48**, 2134–2136 (2012).

[51] Lara Martínez-Fernández, Inés Corral, Giovanni Granucci, and Maurizio Persico, "Competing ultrafast intersystem crossing and internal conversion: a time resolved picture for the deactivation of 6-thioguanine," Chemical Science **5**, 1336–1347 (2014).

[52] Ganglong Cui and Wei-hai Fang, "State-specific heavy-atom effect on intersystem crossing processes in 2-thiothymine: A potential photodynamic therapy photosensitizer," The Journal of chemical physics **138**, 044315 (2013).

[53] Joao Paulo Gobbo and Antonio Carlos Borin, "2-thiouracil deactivation pathways and triplet states population," Computational and Theoretical Chemistry **1040**, 195–201 (2014).

[54] Sebastian Mai, Philipp Marquetand, and Leticia González, "A static picture of the relaxation and intersystem crossing mechanisms of photoexcited 2-thiouracil," The Journal of Physical Chemistry A **119**, 9524–9533 (2015).

[55] Sebastian Mai, Philipp Marquetand, and Leticia González, "Intersystem crossing pathways in the noncanonical nucleobase 2-thiouracil: A time-dependent picture," The journal of physical chemistry letters **7**, 1978–1983 (2016).

[56] Matthias Ruckenbauer, Sebastian Mai, Philipp Marquetand, and Leticia González, "Photoelectron spectra of 2-thiouracil, 4-thiouracil, and 2, 4-dithiouracil," The Journal of chemical physics **144**, 074303 (2016).

[57] Michael Schmidt and Hod Lipson, "Distilling free-form natural laws from experimental data," science **324**, 81–85 (2009).

[58] Roald Hoffmann, "Extended huckel theory. ii. σ orbitals in the azines," The Journal of Chemical Physics **40**, 2745 (1964).

[59] Roald Hoffmann, Akira Imamura, and Warren J Hehre, "Benzynes, dehydroconjugated molecules, and the interaction of orbitals separated by a number of intervening sigma bonds," Journal of the American Chemical Society **90**, 1499–1509 (1968).

[60] Marc H Garner, Mads Koerstz, Jan H Jensen, and Gemma C Solomon, "The bicyclo [2.2. 2] octane motif: A class of saturated group 14 quantum interference based single-molecule insulators," The journal of physical chemistry letters **9**, 6941–6947 (2018).

[61] Jimmy C Kromann, Jan H Jensen, Monika Kruszyk, Mikkel Jessing, and Morten Jørgensen, "Fast and accurate prediction of the regioselectivity of electrophilic aromatic substitution reactions," Chemical science **9**, 660–665 (2018).

[62] Adrian Jinich, Avi Flamholz, Haniu Ren, Sung-Jin Kim, Benjamin Sanchez-Lengeling, Charles AR Cotton, Elad Noor, Alán Aspuru-Guzik, and Arren Bar-Even, "Quantum chemistry reveals thermodynamic principles of redox biochemistry," PLoS computational biology **14**, e1006471 (2018).

[63] Adrian Jinich, Benjamin Sanchez-Lengeling, Haniu Ren, Joshua E Goldford, Elad Noor, Jacob N Sanders, Daniel Segrè, and Alán Aspuru-Guzik, "A thermodynamic atlas of carbon redox chemical space," Proceedings of the National Academy of Sciences **117**, 32910–32918 (2020).

[64] Axel D Becke and Erin R Johnson, "A density-functional model of the dispersion interaction," The Journal of chemical physics **123**, 154101 (2005).

[65] Erin R Johnson and Axel D Becke, "A post-hartree-fock model of intermolecular interactions: Inclusion of higher-order corrections," The Journal of chemical physics **124**, 174104 (2006).

[66] Alberto Otero-de-la Roza, Luc M LeBlanc, and Erin R Johnson, "What is "many-body" dispersion and should i worry about it?" Physical Chemistry Chemical Physics **22**, 8266–8276 (2020).

[67] Christopher R Taylor and Graeme M Day, "Evaluating the energetic driving force for cocrystal formation," Crystal growth & design **18**, 892–904 (2018).

[68] Luc M LeBlanc, Stephen G Dale, Christopher R Taylor, Axel D Becke, Graeme M Day, and Erin R Johnson, "Pervasive delocalisation error causes spurious proton transfer in organic acid–base co-crystals," Angewandte Chemie **130**, 15122–15126 (2018).

[69] Michal Biler, Rory M Crean, Anna K Schweiger, Robert Kourist, and Shina Caroline Lynn Kamerlin, "Ground-state destabilization by active-site hydrophobicity controls the selectivity of a cofactor-free decarboxylase," Journal of the American Chemical Society **142**, 20216–20231 (2020).

[70] Jon Paul Janet and Heather J Kulik, "Resolving transition metal chemical space: Feature selection for machine learning and structure–property relationships," The Journal of Physical Chemistry A **121**, 8939–8954 (2017).

[71] Jon Paul Janet, Fang Liu, Aditya Nandy, Chenru Duan, Tzuhsiung Yang, Sean Lin, and Heather J Kulik, "Designing in the face of uncertainty: Exploiting electronic structure and machine learning models for discovery in inorganic chemistry," Inorganic Chemistry **58**, 10592–10606 (2019).

[72] Jon Paul Janet, Sahasrajit Ramesh, Chenru Duan, and Heather J Kulik, "Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization," ACS central science **6**, 513–524 (2020).

[73] Aditya Nandy, Jiazhou Zhu, Jon Paul Janet, Chenru Duan, Rachel B Getman, and Heather J Kulik, "Machine learning accelerates the discovery of design rules and exceptions in stable metal–oxo intermediate formation," ACS Catalysis **9**, 8243–8255 (2019).

[74] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G Boyd, Yongjin Lee, Berend Smit, and Heather J Kulik, "Understanding the diversity of the metal-organic framework ecosystem," Nature communications **11**, 1–10 (2020).

[75] M Amarouche, G Durand, and JP Malrieu, "Structure and stability of xe+ n clusters," The Journal of chemical physics **88**, 1010–1018 (1988).

[76] Oum Keltoum Kabbaj, Marie-Bernadette Lepetit, and Jean-Paul Malrieu, "Inclusion of dynamical polarization effects is sufficient to obtain reliable energies and structures of he+ n clusters," Chemical physics letters **172**, 483–486 (1990).

[77] Santosh C Gadekar, Vasudevan Dhayalan, Ashim Nandi, Inbal L Zak, Shahar Barkai, Meital Shema Mizrachi, et al., "Rerouting an organocatalytic reaction by intercepting its reactive intermediates," (2020).

[78] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu, "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning," Science **365** (2019).

[79] Jan Hermann, Zeno Schätzle, and Frank Noé, "Deep-neural-network solution of the electronic schrödinger equation," Nature Chemistry **12**, 891–897 (2020).

[80] Jens K Nørskov, Thomas Bligaard, Ashildur Logadottir, S Bahn, Lars B Hansen, Mikkel Bollinger, H Bengaard, Bjørk Hammer, Z Sljivancanin, Manos Mavrikakis, et al., "Universality in heterogeneous catalysis," Journal of catalysis **209**, 275–278 (2002).

[81] Frank Abild-Pedersen, Jeff Greeley, Felix Studt, Jan Rossmeisl, Ture R Munter, Poul Georg Moses, Egill Skulason, Thomas Bligaard, and Jens Kehlet Nørskov, "Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces," Physical review letters **99**, 016105 (2007).

[82] Andrew J Medford, Aleksandra Vojvodic, Jens S Hummelshøj, Johannes Voss, Frank Abild-Pedersen, Felix Studt, Thomas Bligaard, Anders Nilsson, and Jens K Nørskov, "From the sabatier principle to a predictive theory of transition-metal heterogeneous catalysis," Journal of Catalysis **328**, 36–42 (2015).

[83] Jens Kehlet Nørskov, Thomas Bligaard, Jan Rossmeisl, and Claus Hviid Christensen, "Towards the computational design of solid catalysts," Nature chemistry **1**, 37–46 (2009).

[84] Zhi Wei Seh, Jakob Kibsgaard, Colin F Dickens, IB Chorkendorff, Jens K Nørskov, and Thomas F Jaramillo, "Combining theory and experiment in electrocatalysis: Insights into materials design," Science **355** (2017).

[85] Yanming Ma, Mikhail Eremets, Artem R Oganov, Yu Xie, Ivan Trojan, Sergey Medvedev, Andriy O Lyakhov, Mario Valle, and Vitali Prakapenka, "Transparent dense sodium," Nature **458**, 182–185 (2009).

[86] Dmitrii V Semenok, Ivan A Kruglov, Igor A Savkin, Alexander G Kvashnin, and Artem R Oganov, "On distribution of superconductivity in metal hydrides," Current Opinion in Solid State and Materials Science **24**, 100808 (2020).

[87] Di Zhou, Dmitrii V Semenok, Defang Duan, Hui Xie, Wuhao Chen, Xiaoli Huang, Xin Li, Bingbing Liu, Artem R Oganov, and Tian Cui, "Superconducting praseodymium superhydrides," Science advances **6**, eaax6849 (2020).

[88] Di Zhou, Dmitrii V Semenok, Hui Xie, Xiaoli Huang, Defang Duan, Alex Aperis, Peter M Oppeneer, Michele Galasso, Alexey I Kartsev, Alexander G Kvashnin, et al., "High-pressure synthesis of magnetic neodymium polyhydrides," Journal of the American Chemical Society **142**, 2803–2811 (2020).

[89] Chris J Pickard and RJ Needs, "High-pressure phases of silane," Physical review letters **97**, 045504 (2006).

[90] Chris J Pickard and RJ Needs, "Ab initio random structure searching," Journal of Physics: Condensed Matter **23**, 053201 (2011).

[91] Chris J Pickard, Ion Errea, and Mikhail I Eremets, "Superconducting hydrides under pressure," Annual Review of Condensed Matter Physics **11**, 57–76 (2020).

[92] Chris J Pickard and Richard J Needs, "Structure of phase iii of solid hydrogen," Nature Physics **3**, 473–476 (2007).

[93] Paul Loubeyre, Florent Occelli, and Paul Dumas, "Synchrotron infrared spectroscopic evidence of the probable transition to metal hydrogen," Nature **577**, 631–635 (2020).

[94] Ross T Howie, Christophe L Guillaume, Thomas Scheler, Alexander F Goncharov, and Eugene Gregoryanz, "Mixed molecular and atomic phase of dense hydrogen," Physical Review Letters **108**,

125501 (2012).
[95] Chris J Pickard and RJ Needs, "Highly compressed ammonia forms an ionic crystal," Nature materials **7**, 775–779 (2008).
[96] S Ninet, F Datchi, P Dumas, M Mezouar, G Garbarino, A Mafety, CJ Pickard, RJ Needs, and AM Saitta, "Experimental and theoretical evidence for an ionic crystal of ammonia at high pressure," Physical Review B **89**, 174103 (2014).
[97] Chris J Pickard and RJ Needs, "Aluminium at terapascal pressures," Nature materials **9**, 624–627 (2010).
[98] Chris J Pickard and RJ Needs, "Dense low-coordination phases of lithium," Physical review letters **102**, 146401 (2009).
[99] Chris J Pickard and RJ Needs, "Predicted pressure-induced s-band ferromagnetism in alkali metals," Physical review letters **107**, 087201 (2011).
[100] Silvia Amabilino, Lars A Bratholm, Simon J Bennie, Alain C Vaucher, Markus Reiher, and David R Glowacki, "Training neural nets to learn reactive potential energy surfaces using interactive quantum chemistry in virtual reality," The Journal of Physical Chemistry A **123**, 4486–4499 (2019).
[101] Moritz P Haag, Alain C Vaucher, Maël Bosson, Stéphane Redon, and Markus Reiher, "Interactive chemical reactivity exploration," arXiv preprint arXiv:1405.4036 (2014).
[102] Maike Bergeler, Gregor N Simm, Jonny Proppe, and Markus Reiher, "Heuristics-guided exploration of reaction mechanisms," Journal of chemical theory and computation **11**, 5712–5722 (2015).
[103] Gregor N Simm and Markus Reiher, "Context-driven exploration of complex chemical reaction networks," Journal of chemical theory and computation **13**, 6108–6119 (2017).
[104] Lise Brethous, Noemi Garcia-Delgado, Julian Schwartz, Sonia Bertrand, Daniel Bertrand, and Jean-Louis Reymond, "Synthesis and nicotinic receptor activity of chemical space analogues of n-(3 r)-1-azabicyclo [2.2. 2] oct-3-yl-4-chlorobenzamide (pnu-282,987) and 1, 4-diazabicyclo [3.2. 2] nonane-4-carboxylic acid 4-bromophenyl ester (ssr180711)," Journal of medicinal chemistry **55**, 4605–4618 (2012).
[105] Ivan Di Bonaventura, Stéphane Baeriswyl, Alice Capecchi, Bee-Ha Gan, Xian Jin, Thissa N Siriwardena, Runze He, Thilo Köhler, Arianna Pompilio, Giovanni Di Bonaventura, et al., "An antimicrobial bicyclic peptide from chemical space against multidrug resistant gram-negative bacteria," Chemical communications **54**, 5130–5133 (2018).
[106] Stefano Sanvito, Corey Oses, Junkai Xue, Anurag Tiwari, Mario Zic, Thomas Archer, Pelin Tozman, Munuswamy Venkatesan, Michael Coey, and Stefano Curtarolo, "Accelerated discovery of new magnets in the heusler alloy family," Science advances **3**, e1602241 (2017).
[107] Alessandro Lunghi and Stefano Sanvito, "Surfing multiple conformation-property landscapes via machine learning: Designing single-ion magnetic anisotropy," The Journal of Physical Chemistry C **124**, 5802–5806 (2020).
[108] Mads C Nielsen, Eirik Lyngvi, and Franziska Schoenebeck, "Chemoselectivity in the reductive elimination from high oxidation state palladium complexes–scrambling mechanism uncovered," Journal of the American Chemical Society **135**, 1978–1985 (2013).
[109] Mads C Nielsen, Karl J Bonney, and Franziska Schoenebeck, "Computational ligand design for the reductive elimination of arcf3 from a small bite angle pdii complex: remarkable effect of a perfluoroalkyl phosphine," Angewandte Chemie International Edition **53**, 5903–5906 (2014).
[110] Maoping Pu, Italo A Sanhueza, Erdem Senol, and Franziska Schoenebeck, "Divergent reactivity of stannane and silane in the trifluoromethylation of pdii: cyclic transition state versus difluorocarbene release," Angewandte Chemie International Edition **57**, 15081–15085 (2018).
[111] Peng Bai, Mi Young Jeon, Limin Ren, Chris Knight, Michael W Deem, Michael Tsapatsis, and J Ilja Siepmann, "Discovery of optimal zeolites for challenging separations and chemical transformations using predictive materials modeling," Nature communications **6**, 1–9 (2015).
[112] Alexander J Sodt, Michael Logan Sandar, Klaus Gawrisch, Richard W Pastor, and Edward Lyman, "The molecular structure of the liquid-ordered phase of lipid bilayers," Journal of the American Chemical Society **136**, 725–732 (2014).
[113] Stephen Whitelam and Isaac Tamblyn, "Learning to grow: Control of material self-assembly using evolutionary reinforcement learning," Physical Review E **101**, 052604 (2020).
[114] Yinan Shu and Donald G Truhlar, "Diabatization by machine intelligence," Journal of Chemical Theory and Computation **16**, 6456–6464 (2020).
[115] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space," New Journal of Physics **15**, 095003 (2013).
[116] O Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko, "Exploring chemical compound space with quantum-based machine learning," Nature Reviews Chemistry **4**, 347–358 (2020).
[117] Masato Sumita, Xiufeng Yang, Shinsuke Ishihara, Ryo Tamura, and Koji Tsuda, "Hunting for organic molecules with artificial intelligence: molecules optimized for desired excitation energies," ACS central science **4**, 1126–1133 (2018).
[118] Fiorella Ruggiu, Patrick Gizzi, Jean-Luc Galzi, Marcel Hibert, Jacques Haiech, Igor Baskin, Dragos Horvath, Gilles Marcou, and Alexandre Varnek, "Quantitative structure–property relationship modeling: a valuable support in high-throughput screening quality control," Analytical chemistry **86**, 2510–2520 (2014).
[119] Peter B Jensen, Agata Bialy, Didier Blanchard, Steen Lysgaard, Alexander K Reumert, Ulrich J Quaade, and Tejs Vegge, "Accelerated dft-based design of materials for ammonia storage," Chemistry of Materials **27**, 4552–4561 (2015).
[120] Vladimir Tripkovic, Heine Anton Hansen, and Tejs

Vegge, "From 3d to 2d co and ni oxyhydroxide catalysts: Elucidation of the active site and influence of doping on the oxygen evolution activity," ACS Catalysis **7**, 8558–8571 (2017).

[121] Jinsong Wang, Jia Liu, Bao Zhang, Feng Cheng, Yunjun Ruan, Xiao Ji, Kui Xu, Chi Chen, Ling Miao, and Jianjun Jiang, "Stabilizing the oxygen vacancies and promoting water-oxidation kinetics in cobalt oxides by lower valence-state doping," Nano Energy **53**, 144–151 (2018).

[122] Bing Huang and O Anatole Von Lilienfeld, "Understanding molecular representations in machine learning: The role of uniqueness and target similarity," The Journal of Chemical Physics **145**, 161102 (2016).

[123] Felix A Faber, Anders S Christensen, Bing Huang, and O Anatole Von Lilienfeld, "Alchemical and structural distribution based representation for universal quantum machine learning," The Journal of chemical physics **148**, 241717 (2018).

[124] Felix A Faber, Alexander Lindmaa, O Anatole Von Lilienfeld, and Rickard Armiento, "Machine learning energies of 2 million elpasolite (a b c 2 d 6) crystals," Physical review letters **117**, 135502 (2016).

[125] Andrew Shamp and Eva Zurek, "Superconductivity in hydrides doped with main group elements under pressure," Novel Superconducting Materials **3**, 14–22 (2017).

[126] Eva Zurek and Tiange Bi, "High-temperature superconductivity in alkaline and rare earth polyhydrides at high pressure: A theoretical perspective," The Journal of chemical physics **150**, 050901 (2019).

[127] Wenwen Cui, Tiange Bi, Jingming Shi, Yinwei Li, Hanyu Liu, Eva Zurek, and Russell J Hemley, "Route to high-t c superconductivity via ch 4-intercalated h 3 s hydride perovskites," Physical Review B **101**, 134504 (2020).