

Alchemical normal modes unify chemical space

Stijn Fias,^{*,†} K. Y. Samuel Chang,[¶] and O. Anatole von Lilienfeld^{*,¶}

[†]*General Chemistry (ALGC), Vrije Universiteit Brussel (Free University Brussels - VUB),
Pleinlaan 2, 1050 Brussel, Belgium*

[‡]*Department of Chemistry & Chemical Biology, McMaster University, Hamilton, ON,
Canada L8S 4L8*

[¶]*Institute of Physical Chemistry and National Center for Computational Design and
Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel,
4056 Basel, Switzerland*

E-mail: fias@mcmaster.ca; anatole.vonlilienfeld@unibas.ch

Abstract

In silico design of new molecules and materials with desirable quantum properties by high-throughput screening is a major challenge due to the high dimensionality of chemical space. To facilitate its navigation, we present a unification of coordinate and composition space in terms of alchemical normal modes (ANMs) which result from second order perturbation theory. ANMs assume a predominantly smooth nature of chemical space and form a basis in which new compounds can be expanded and identified. We showcase the use of ANMs for the energetics of the iso-electronic series of diatomics with 14 electrons, BN doped benzene derivatives ($C_{6-2x}(BN)_xH_6$ with $x = 0, 1, 2, 3$), predictions for over 1.8 million BN doped coronene derivatives, and genetic energy optimizations in the entire BN doped coronene space. Using Ge lattice scans as reference, the applicability ANMs across the periodic table is demonstrated for III-V and IV-IV-semiconductors Si, Sn, SiGe, SnGe, SiSn, as well as AlP, AlAs, AlSb, GaP, GaAs, GaSb, InP, InAs, and InSb. Analysis of our results indicates simple qualitative structure property rules for estimating energetic rankings among isomers. Useful quantitative estimates can also

be obtained when few atoms are changed to neighboring or lower lying elements in the periodic table. The quality of the predictions often increases with the symmetry of system chosen as reference due to cancellation of odd order terms. Rooted in perturbation theory the ANM approach promises to generally enable unbiased compound exploration campaigns at reduced computational cost.

1 Introduction

A quantum mechanics based understanding of chemical compound space (CCS) is crucial for gauging the predictive power and versatility of theoretical chemistry models, as well as and for the computational design of molecular and solid matter. Due to its universality to account for the physics of electrons which govern the behavior of matter the use of quantum mechanics is mandatory in this context. The complexity of its solutions, however, hampers the intuitive understanding and conceptualization of the solutions obtained. High-throughput-screening campaigns have therefore been proposed to tackle materials design challenges,¹⁻³ and extensive materials quantum data records have been established.⁴⁻⁷ Still, the high dimensionality of CCS^{8,9} combined with the consid-

erable cost for repeatedly evaluating quantum properties from scratch severely hampers even the most sophisticated optimization algorithms, let alone screening.

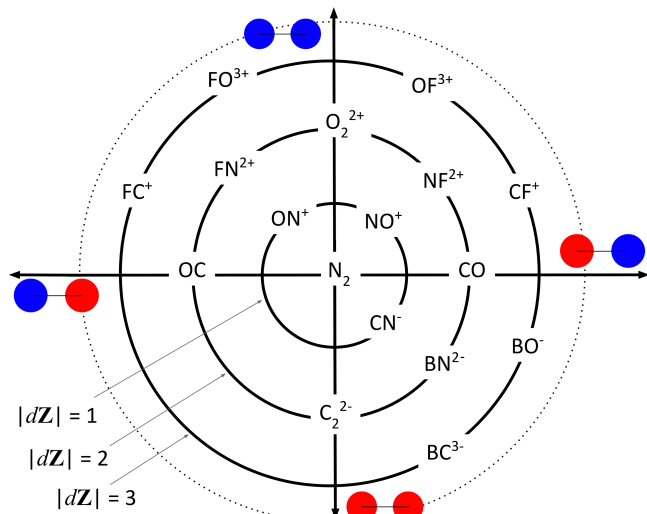


Figure 1: The shell-structure of alchemical hyper-spheres, illustrated for the chemical space of all diatomics with 14 electrons ($\text{ANM}_1 = 1/\sqrt{2} - 1/\sqrt{2}$, $\text{ANM}_2 = 1/\sqrt{2} + 1/\sqrt{2}$, $\epsilon_1 = -3.65$ a.u., $\epsilon_2 = -2.987$ a.u.). Homo-nuclear diatomics (vertical axis) correspond to the ridge. Interatomic distance dependence is shown in 2 for all neutral diatomics (horizontal axis).

Given thousands of previously acquired representative reference examples used for training, quantum machine learning models have recently emerged as a viable option to further accelerate materials design by multiple orders of magnitude,^{10–15} reaching prediction errors on par with DFT.^{16,17} Alas, also these methods require representative training sets, and the combinatorial nature of chemistry simply prohibits the establishment of a comprehensive encyclopedia. Consequently, more powerful approaches are needed, e.g. exploiting a more rigorous notion of chemical space.¹⁸ Here, we investigate such an alternative, physics- rather than statistics-based approach for the sampling of CCS which reaches machine learning speed and accuracy. Instead of massive training sets which must be representative, it requires only a single reference calculation which must be relevant. It is rooted in second-order perturbation theory and includes variations in nuclear

charges, a.k.a. “alchemical changes”. Alchemical perturbations have been used in quantum chemistry every since Hückel’s work on predicting substituent effects in benzene,¹⁹ and Pauling’s follow up work.^{20,21} More modern approaches include Refs.,^{22–31} and more recently, substantial progress has been made along similar lines^{32–41} using first and second order perturbation theory to introduce alchemical normal modes (ANMs), resulting from diagonalization of a unified Hessian, to form a complete, low-dimensional, and intuitive basis which spans CCS. Building on this, we provide a novel understanding of the structure of chemical space, and we show how to utilize it for solving inverse design problems with unprecedented speed and accuracy. The expansion of individual query molecules in their ANMs enables rapid energy estimates which we demonstrate for screening over 1.8 M BN-doped coronene derivatives based on a single quantum reference calculation.

The remainder of this paper is structured as follows, we introduce the theoretical underpinnings of the ANMs in the Theory section, exemplifying their usage for molecular nitrogen which is sufficiently simple to easily gain an intuition. Subsequently, we demonstrate and assess the performance of ANMs for the complete CCS of all BN doped benzene derivatives. ANMs of coronene are then used to (i) predict electronic energies of over 1.8 M of its BN doped derivatives, and (ii) to discover those structures with lowest and highest lying energy, as identified by a genetic optimization algorithm. The applicability of ANMs is also demonstrated for solid systems, as exemplified for III-V and IV semi-conductors. After the discussion of our results we briefly conclude this investigation. Finally, methodological details are given for the computational aspects.

2 Theory

Within the Born-Oppenheimer approximation, the total potential ground-state energy of a compound, $U = E + V_{NN}$, consists

of the nuclear Coulomb repulsion ($V_{NN} = \sum_{I>J} Z_I Z_J / |\mathbf{R}_I - \mathbf{R}_J|$) and the electronic energy E , the solution of the electronic Schrödinger equation (SE). In order to facilitate the discussion, all results and discussions in the following will be concerned exclusively with the latter. Subsequent addition of the V_{NN} -term, often necessary when aiming for comparison to experimental numbers, is trivial since composition and coordinates are always assumed to be known. From the quantum mechanical point of view of the potential energy hypersurface, systems differ only by nuclear charges $\{Z_I\}$, atomic coordinates $\{\mathbf{R}_I\}$, and number of electrons N . Within second order perturbation theory, we can therefore Taylor expand the electronic energy of any target system \mathbf{x}^t around the electronic energy of a reference system \mathbf{x}_0 ,

$$E(\mathbf{x}^t) = E(\mathbf{x}_0) + \mathbf{g}d\mathbf{x} + \frac{1}{2}d\mathbf{x}^T \mathbf{H}d\mathbf{x} + \dots (1)$$

where $\mathbf{x} = (Z_1, Z_2, \dots, Z_M, \mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_M, N)$, and \mathbf{g} and \mathbf{H} represent a unified gradient and Hessian, respectively. First order terms are firmly established for all variables through the Hellmann-Feynman theorem for changes in nuclear positions (to relax or run ab initio molecular dynamics⁴²), and charges.^{22,27,32,33,36,37,43-45} The derivative with respect to N is related to ionization potential and electron affinity by virtue of Koopman's and Janak's theorem,⁴⁶ and exhibits the well established derivative discontinuity at integer N ,^{47,48} so important for the construction of improved exchange-correlation approximations.⁴⁹

Some elements in the Hessian,

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 E_0}{\partial Z_I \partial Z_J} & \frac{\partial^2 E_0}{\partial Z_I \partial \mathbf{R}_J} & \frac{\partial^2 E_0}{\partial Z_I \partial N} \\ \frac{\partial^2 E_0}{\partial \mathbf{R}_I \partial Z_J} & \frac{\partial^2 E_0}{\partial \mathbf{R}_I \partial \mathbf{R}_J} & \frac{\partial^2 E_0}{\partial \mathbf{R}_I \partial N} \\ \frac{\partial^2 E_0}{\partial N \partial Z_J} & \frac{\partial^2 E_0}{\partial N \partial \mathbf{R}_J} & \frac{\partial^2 E_0}{\partial N^2} \end{bmatrix} \quad (2)$$

are also part of text-book chemistry: The coordinate subspace matrix corresponds to the conventional Hessian, related to the harmonic

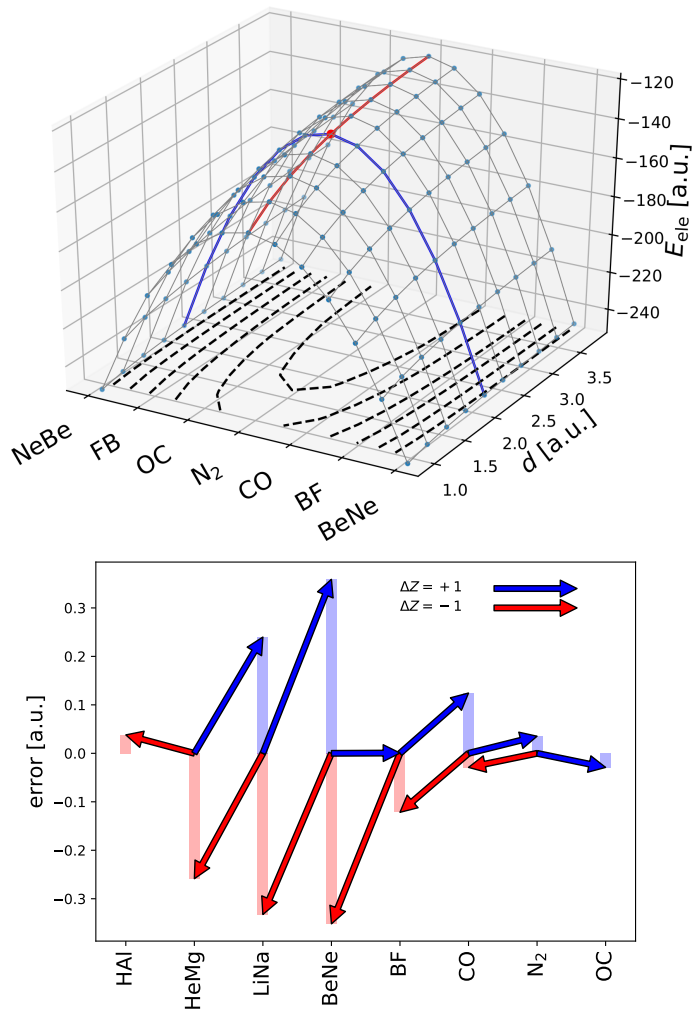


Figure 2: TOP: The electronic energy is shown as a function of interatomic distance and difference in nuclear charge. BOTTOM: The error of ANM based predictions of neighboring systems is shown at fixed interatomic distance at 1.1 Å.

molecular vibrational normal modes, or the second order derivative of the electronic energy with respect to the number of electrons is the chemical hardness, introduced by Parr and Pearson.⁵⁰ The $\frac{\partial^2 E_0}{\partial Z_I \partial Z_J}$ block corresponds to the alchemical hardness.^{35,43} The least conventional off-diagonal blocks correspond to nuclear Fukui functions, $\frac{\partial^2 E_0}{\partial N \partial \mathbf{R}_J}$,⁵¹ alchemical Fukui Functions $\frac{\partial^2 E_0}{\partial N \partial Z_J}$,^{43,44} and the alchemical force, $\frac{\partial^2 E_0}{\partial \mathbf{R}_I \partial Z_J} = \int d\mathbf{r}(\partial_{\mathbf{R}_I} \rho(\mathbf{r})) / |\mathbf{r} - \mathbf{R}_J| = \int d\mathbf{r}(\rho(\mathbf{r}) + Z_J \partial_{Z_J} \rho(\mathbf{r})) (\mathbf{r} - \mathbf{R}_I) / |\mathbf{r} - \mathbf{R}_I|^3$ (due to Maxwell-relation). To the best of our knowledge, such a unified Hessian has not yet been studied in full, despite the well-known non-linearities of quantum properties in chemical space.

In analogy to vibrational normal modes, diagonalization of this unified Hessian for any meaningful reference system defines an orthogonal and complete basis in which other chemical compounds and their intra-molecular motion can be expanded. The resulting eigenvalues and eigenfunctions, the ‘‘alchemical normal modes’’ (ANMs), correspond to principal curvatures, and thus carry fundamental importance for our understanding of CCS (Transformation from Hessian matrix to second fundamental form may be required). The composition of any target molecule can be linearly expanded in the complete vector basis spanned by the ANM matrix (\mathbf{Q}) of the reference compound \mathbf{x}_0 , i.e. $\mathbf{x}^t = \mathbf{x}_0 + d\mathbf{x}$. The necessary coefficients are given by $\mathbf{c} = \mathbf{Q}d\mathbf{x}$, resulting in the second order energy estimate,

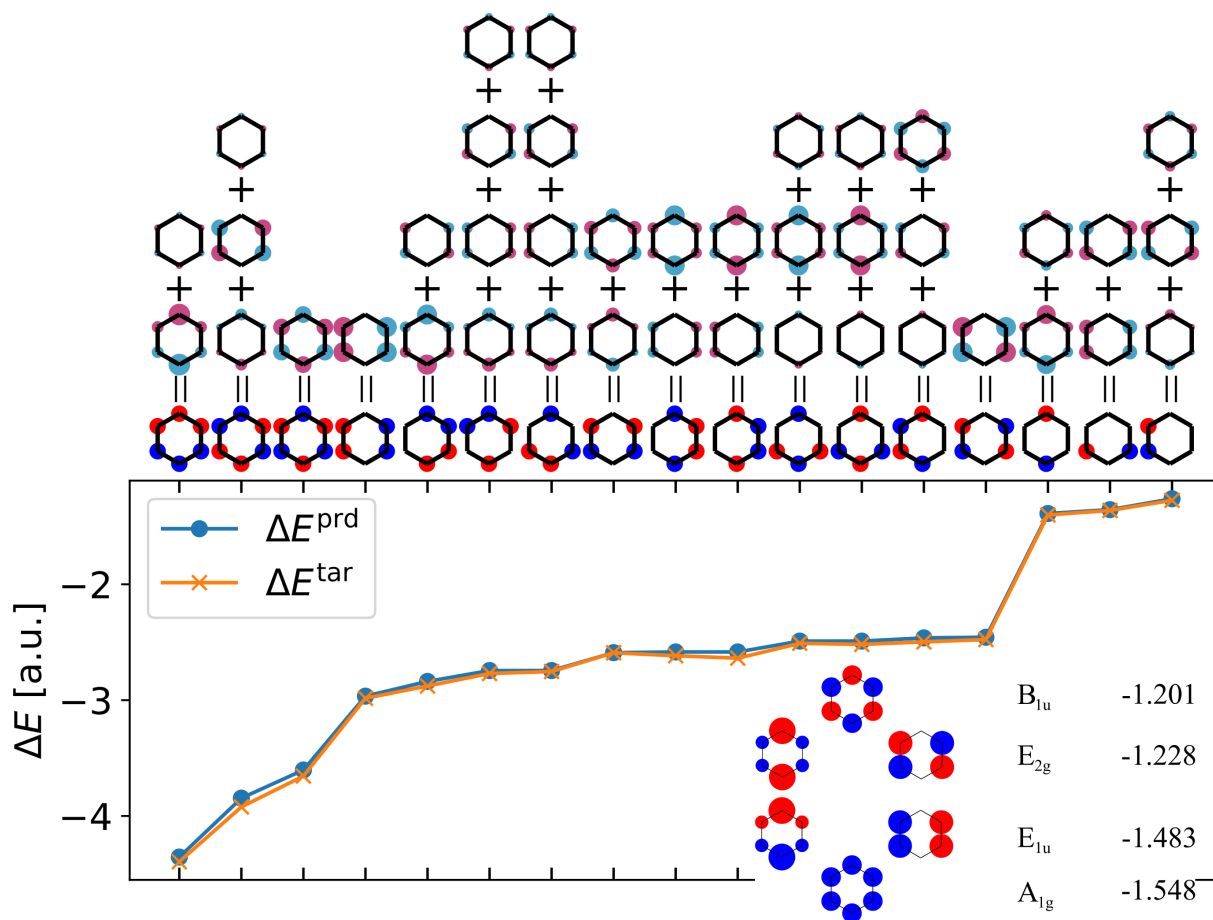
$$E(\mathbf{x}^t) \approx E(\mathbf{x}_0) + \mathbf{g}d\mathbf{x} + \frac{1}{2} \mathbf{c}^T \mathbf{Y} \mathbf{c} \quad (3)$$

where \mathbf{Y} is the diagonal eigenvalue matrix of the unified Hessian \mathbf{H} . This framework leads to an encompassing definition of the structure of CCS which couples configurational, compositional, and electronic degrees of freedom. Let us consider projections onto lower dimensional manifolds of this structure. Firstly, for fixed composition ($|\{Z_I\}|$), the conventional picture of changes in configurations (geometry) and electron number (redox-properties), emerges. Secondly, when fixing geometry and electron number for n atom systems instead, an

n -dimensional alchemical hyper-sphere (AHS) can be defined for reference compounds with maximal symmetry (*vide infra* why) being at the origin (i.e. that system for which all atoms have same nuclear charge, $Z_I = N_p/n \forall I$ where $N_p = |\mathbf{Z}|$). The AHS has a shell structure where integer nuclear charge combinations emerge for integer radii, $|d\mathbf{Z}|$, i.e. systems with a correspondence in reality. Fig. 1 illustrates the AHS for the di-atomics with $N = 14$ electrons which can be expanded in ANMs of molecular nitrogen. ANM q_1 corresponds to charge-neutral simultaneous depletion and growth of the nuclear charge at the two respective atomic sites, covering the series N_2 , CO , BF , ..., AlH , Si . ANM q_2 corresponds to the simultaneous addition or removal of protons at the two respective atomic sites, covering the series ..., B_2^{4-} , C_2^{2-} , N_2 , O_2^{2+} , F_2^{4+} , ... Linear combinations of q_1 and q_2 define all the other possible diatomics which can be defined on shells with radii $|d\mathbf{Z}| = 1, 2, 3, \dots$ e.g. NO^+ , expanded in ANMs of N_2 , corresponds to $1/\sqrt{2}q_1 + 1/\sqrt{2}q_2$. Obviously, while target compounds with large q_2 component will be increasingly charged and unstable without changes in electron number, in the absence of external fields or extreme conditions, this extended unified structure of CCS is general in scope as it accounts for a continuum of ‘‘alchemical’’ chemistries with fractional nuclear charges. We note that extensions of reality to include such fictitious degrees of freedom have a long-standing track-record in thermodynamics and statistical mechanics, e.g. in the form of extended Lagrangians, and can be used for any state function.

3 Results and discussion

Within the first subsection we present and discuss results obtained for the neutral iso-electronic diatomic series with 14 electrons, including all interatomic distances and all possible nuclear charge combinations. In order to facilitate the discussion and visualization of results in the subsequent subsections, we restrict ourselves to fixed geometries and electron number, and we focus on changes in composi-



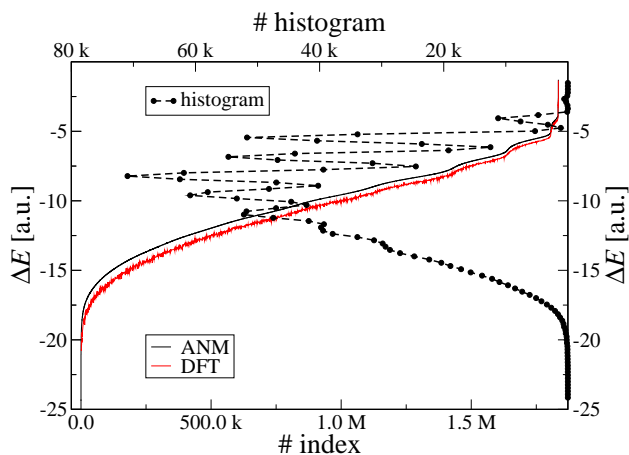
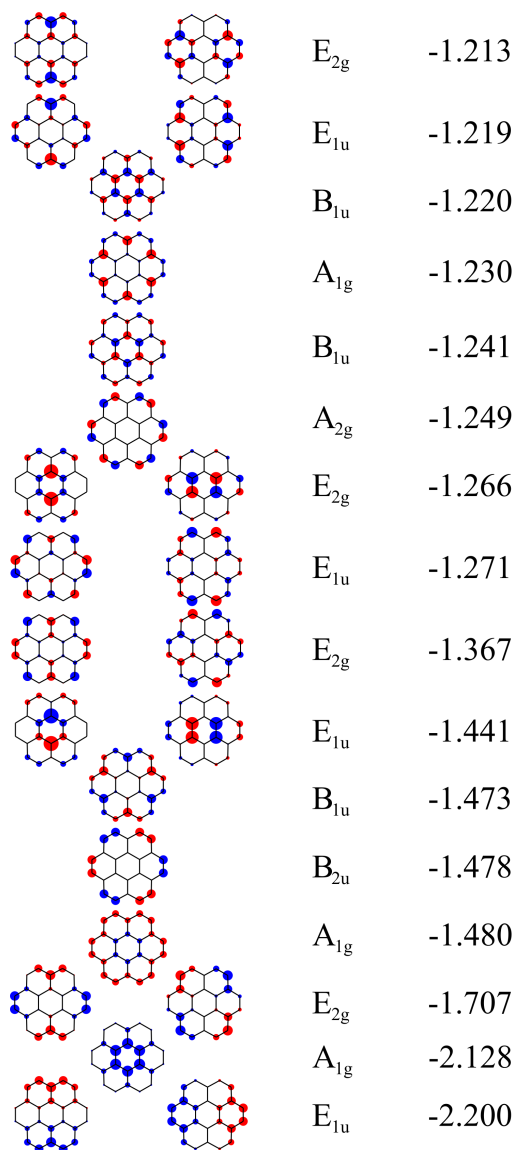


Figure 4: TOP: Alchemical normal modes of coronene and corresponding eigenvalues in Hartree. BOTTOM: ANM based estimates of energy change from coronene for ~ 1.8 M BN doped coronene mutants in ascending order (black solid). Their distribution is shown as a histogram (black dashed). Validating DFT re-

tion only without any loss of generality. This restriction is obviously severe for large, high-dimensional systems which sample many effective degrees of freedom, e.g. proteins with many shallow conformational minima. However, for materials classes with rigid lattices and an effectively low dimensionality e.g. crystals with high symmetry under ambient conditions, the relevant configurational degrees of freedom can easily be scanned and enable the exploration of combinatorially growing compositional spaces with ease.

3.1 Diatomic series with 14 electrons

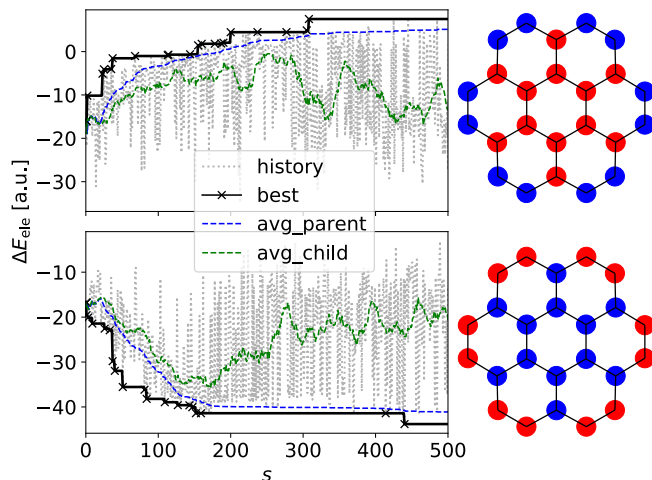


Figure 5: Genetic algorithm maximization (upper panel) and minimization (lower panel) on electronic energy for $(\text{BN})_{12}\text{H}_{12}$. The difference electronic energy for optimization history (gray dotted lines), current optimal molecule (black crosses), average over the parent pool (blue dashed lines) and the corresponding children (green dashed lines) are plotted respectively. The optimized molecules are shown on the right where B and N are represented by red and blue atoms.

The projection of the unified ANMs based structure of CCS onto lower dimensional realistic systems is straight-forward. We exemplify this for the electronic energy of the neutral iso-electronic $N = 14$ electron series with

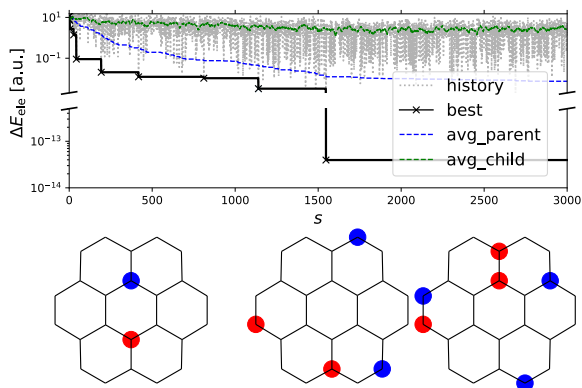


Figure 6: Genetic algorithm energy optimization run. The target energy corresponds to left hand molecule. The black line (best) corresponds to the best molecule in population. The blue line (average) corresponds to the average value of the population. The target molecule is discovered within 7 steps out of ~ 4 billion compounds. Energies of mid and right hand side molecule deviate only by 0.4 and 0.5 mHa from target.

variable \mathbf{Z} and $\{\mathbf{R}_I\}$, as shown in Fig. 2. Note the text-book dependence of the electronic energy on interatomic distance (decaying towards the united atom energy of Si), the well known concavity for fixed N and $\{\mathbf{R}\}$,¹⁸ and the ridge corresponding to the homo-diatomic N_2 ($\Delta Z = 0$). The apparent monotonic and smooth behavior of the electronic energy in this sub-space corroborates the applicability of perturbation theory. This implies that, in analogy to vibrational normal modes, the gradient with respect to Z must be zero at the ridge. And it is obvious, indeed, that the corresponding Hellmann-Feynman derivative, $\int d\mathbf{r} \rho(\mathbf{r}) \left(\frac{1}{|\mathbf{r}-\mathbf{R}_1} - \frac{1}{|\mathbf{r}-\mathbf{R}_2} \right)$,³⁰ must be zero due to the symmetry of the electron density, just as well as all higher odd order energy derivatives. As such, when mutating nuclear charges, the reference system with symmetrical atomic densities will always correspond to a maximum in the electronic energy E . This observation would suggest that it is preferable to select reference systems with maximal symmetry in order to quench odd higher order effects.

Numerical electronic energy estimates of alchemically adjacent systems $|d\mathbf{Z}| = 2$ within the same neutral iso-electronic diatomic series with $N = 14$ electrons (lower panel in Fig. 2) support this idea: The prediction error increases systematically for estimates of CO, BF, and BeNe when using systems as reference which decrease in symmetry, i.e. N_2 , CO, and BF, respectively. These results confirm, not surprisingly, that the harmonic approximation works best at the ridge, in complete analogy to harmonic vibrational normal-modes working best at zero Kelvin. The error becomes largest for changes involving substantial changes in electron densities, e.g. when valence electrons flow from p to s orbitals and from principal quantum number 2 to 1 and 3 (for example when targeting or referencing BeNe, LiNa, or HeMg). Also note the negative sign of the error as one predicts CO from N_2 , BF from CO, and BeNe from BF. This implies an exponent of the actual energy surface which is larger than 2, which is in line with independent findings for the energy of free atoms scaling as $\sim -Z^{7/3}$.¹ It is also interesting to note the left/right anti-symmetry in the error of most predictions, e.g. the error made when predicting the electronic energy of CO using BF as a reference has the same magnitude as its reverse counterpart, i.e. predicting the electronic energy of BF using CO as a reference. This suggests, that the exponent is not much larger than 2. And it is to be contrasted with the findings for first order based estimates of energy changes, e.g. for alchemical predictions of covalent bond energies,^{34,35} where, due to the concavity of the electronic energy in Z , the error is clearly not symmetrical upon exchange of reference and target system.

3.2 BN doping of benzene

While any iso-electronic diatomic series can be expanded in the ANMs of the corresponding homo-diatomic, the maximum ridge in the electronic energy, ANMs become less obvious for larger molecules. Based on above symmetry

¹K. Burke, oral contribution, IPAM reunion 2018

arguments, the benzene molecule with point group D_{6h} emerges as an intuitive 2D polyatomic reference system. Considering all the possible neutral iso-electronic changes of carbon to B and N it is clear that some odd order energy derivatives will be zero due to symmetry. Scaling up the coordinates will lead to the electronic energy of the dissociated free atoms, while scaling them down leads to nuclear fusion, i.e. the energy of the united atom, Mo ($Z = 42$). Here, we remind the reader that we do not consider nuclear Coulomb repulsion, and that for this and the remaining examples, ANM based predictions are always exemplified for changes in nuclear charges only, i.e. keeping coordinates and electron numbers fixed.

Fig. 3 illustrates eigenvalues and ANMs of benzene and their use for predicting the iso-electronic doping of benzene with B and N (keeping number of electrons and geometry constant) at all possible atomic sites, i.e. for all possible constitutional isomers with sum formula C_4BNH_6 , $C_2(BN)_2H_6$, and $(BN)_3H_6$. It is intriguing to note the similarity of form, degeneracy, and energy ordering to ordinary Hückel orbitals of benzene. Also note that, in analogy to Hückel, the eigenvalues decrease as the number of nodes in the ANM decrease. Obviously, however, there is no π -electron structure at the atom’s origin, and also the eigenvalues do not correspond to solutions to Hückel’s secular equation. While the eigenvalues depend on the level of theory used, because of symmetry the ANMs are independent of that.

Apart from their appealingly simple and insightful structure, one can use these ANMs to easily estimate relative energetics of possible mutations on the back of an envelope. For example, using ANM based CCSD predictions of the electronic energies according to Eq. 3, i.e. $E(\mathbf{x}^t) \approx \sum_i \epsilon_i c_i^2$, of the three constitutional isomers of BN doped benzene result in -1.226, -1.312, -1.348 Ha for ortho, meta, and para substitutions, respectively. Compared to actual values, the energetic ordering is conserved, and the estimates are in decent agreement with the corresponding CCSD energies (-1.275, -1.362, -1.401 Ha), i.e. systematically overestimating the truth by ~ 0.05 Ha. Predicted and actual

changes in energy with respect to pure benzene are also on display in Fig. 3 for all the possible mutants, and indicate very decent qualitative agreement. Qualitatively, the energetic order can also be explained by noting that the closer the poles of the perturbing potential, the smaller the integral of their product with the electron density response, the smaller the deviation from the energy of benzene. This is consistent with the fact that ANMs with fewer nodes have lower eigenvalues. By consequence, the isomers with sum formula $(BN)_3H_6$ will decrease in energy when decreasing the number of nodal surfaces between B and N mutations, i.e. $E(B_3N_3H_6) < E(B_2NB_2H_6) < E(BNBNBH_6)$. Inspection of the linear combination of ANMs resulting in each of these isomers also clearly indicates that the energy decays as ANMs with fewer nodes are being blended in. Unfortunately, before one can compare to experiments, addition of the nuclear repulsion terms will obfuscate this ranking unless the inequalities introduced by Mezey can be applied.²⁴ We believe nevertheless that these rules are obviously useful for estimating the ranking of electronic energies in constitutional isomers which is of utmost relevance for gaining a deepened and more intuitive grasp of quantum chemistry based relationships. We do not think that these rules have been noted yet.

3.3 BN doping of coronene

In order to explore the applicability of this approach to the computational design problem of real materials we have considered the case of BN doping also for coronene ($C_{24}H_{12}$), relevant for molecular electronics applications.^{52,53} BN doping of coronene results in the 24 ANMs and eigenvalues shown in the upper panel of Fig. 4. Coronene possesses three symmetrically distinct classes of carbon atoms: Six atoms corresponding to the inner ring, six atoms bonded to the inner ring, and the twelve outer atoms. Consequently, first order derivatives with respect to iso-electronic BN doping within any of these three groups of atoms are zero. Doping with BN pairs within these classes, we have generated over 1.8 M mutants, and estimated their

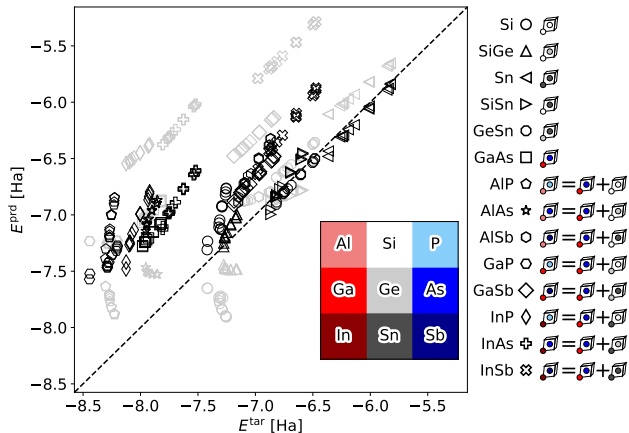


Figure 7: Alchemically predicted vs. true scatter plot for first (gray) and second order (black) estimates of total energies for various III-V and IV-IV semiconductors expanded in Ge ANMs in periodic table (blue/red within period, white/black within column).

energy based on coronene’s ANMs. Duplicates have been removed using the sorted Coulomb matrix representation.^{10,54} The resulting energies are shown in ascending order in the lower panel of Fig. 4, together with a sub-sample of more than two thousand validating cases for which the corresponding DFT PBE energy has been calculated. Clearly, the overall qualitative trends of DFT and ANM based estimates agree well with each other. A linear fit of predictions to validations for the 2 k mutants yields a MAE of ~ 28 kcal/mol. Further analysis indicates that the error grows with number of BN pairs by, on average, ~ 6.5 kcal/mol per BN pair. In the case of the hundred coronene mutants closest in energy to coronene, for example, the MAE amounts to only ~ 2.1 kcal/mol. Fig. 4 also reports the energy distribution of the 1.8 M coronene mutants. Discrete peaks correspond to higher lying mutant stoichiometries with less BN content.

We have also explored the usefulness of ANM based energy estimates for the molecular design challenge of finding those constitutional isomers of $(\text{BN})_{12}\text{H}_{12}$ with the respectively lowest and highest electronic energy. A genetic optimization algorithm based on first (which now can be non-zero due to BN doping among symmetrically inequivalent carbon atoms) and second

order energy estimates only converges within a few hundred steps. The optimization history, together with the converged molecules, are on display in Fig. 5. The most and least stable isomers correspond, not surprisingly, to those N and B distributions which localize the valence electron density most and least, respectively. Results in Fig. 6 summarize the genetic optimization history when searching for the energy of the coronene mutant resulting from BN doping in para position of the inner carbon ring. The target molecule is identified by the genetic optimizer after just over 1500 optimization steps.

These calculations can serve to illustrate the scope of the computational savings which result from the use of ANMs: ~ 1000 CPU core hours were necessary on average to calculate the over 2 k validating DFT energies. The ANM based estimates of 1.8 M mutants, by comparison, incurred negligible overhead (~ 30 CPU core minutes).

3.4 Expanding III-V and IV-IV semiconductors in ANMs of Ge

Finally, we have investigated the applicability of ANMs to solids. More specifically, we have considered iso-valence-electronic expansions in the ANMs resulting from a minimal unit cell in the pseudopotential parameter space of two Ge atoms. Using 15 parameters in the analytical pseudopotentials of Goedecker and Hutter,^{55,56} four dimensions have been considered per atom: Right and left in a period of the periodic table (analogous to Z as discussed above), and up and down in a column of the periodic table (corresponding to changes in principal quantum number). The pseudopotential parameters were coupled to these dimensions with the chain-rule, as explained in the Methods section. The resulting projection yields the ANMs of Ge used for expansion as shown in Fig. 7 for all possible IV-IV and III-V semiconductors which neighbor Ge in the periodic table, i.e. Si, Sn, SiGe, SnGe, SiSn, as well as AlP, AlAs, AlSb, GaP, GaAs, GaSb, InP, InAs, and InSb. More specif-

ically, the figure shows first and second order total potential energy estimates vs. actual DFT evaluations for various lattice scans. Note that first order gradients are non-zero due to lack of symmetry in pseudopotential parameter space, i.e. while the reference electron density is symmetric, the perturbing potential is not perfectly anti-symmetric. A clear correlation is found for first order estimates. Inclusion of second order contributions through ANM based predictions improves the overall correlation, and, maybe more importantly, results in a systematic overestimation of the energies of the target systems (consistent with aforementioned observations made for molecules). The prediction quality for IV-IV crystals is particularly encouraging, in all likelihood profiting from near-linear changes in valence electron density as one changes from one period to the next. ANM based estimates of III-V materials, however, are more challenging. However, the errors appear to be rather systematic in their overestimation. This raises hope that it can still be quite feasible to correct it once third order contributions are being included.

4 Conclusions

The Born-Oppenheimer approximation implies a parametric dependency of the electronic ground-state energy on nuclear positions, charges, and electron number. In order to obtain a general yet rigorous framework of chemical space, we have unified all the relevant degrees of freedom by extending the ordinary normal mode procedure used for atomic positions by alchemical normal modes (ANMs) which also include nuclear charges and electron number. Applied within Taylor expansions, the energy of iso-electronic target compounds can be expanded in ANMs. The resulting estimates are exact up to third order if electron densities in reference system and perturbing potential are symmetric and anti-symmetric, respectively. We have illustrated the concept for diatomics using molecular nitrogen as a reference, for all and ~ 1.8 M BN doped mutants of benzene and coronene, respectively. The ap-

plicability to solids has been demonstrated for all III-V and IV-IV semiconductors neighboring GeGe. Future extension to higher orders to improve predictive power, to other properties, excited states, and generalizations to entire functional groups can also be envisioned.

5 Methods

6 Unified Hessian matrix for N_2

The full Hessian matrix as defined in Eq. 2 for N_2 at its equilibrium geometry (using PBE⁵⁷ with uncontracted cc-pVDZ basis function in Gaussian09⁵⁸). All matrix elements are rounded to the third decimal numbers.

$$\mathbf{H} = \begin{pmatrix} \Delta Z_1 & \Delta Z_2 & \Delta \mathbf{R} & \Delta N \\ -3.126 & 0.139 & 0.121 & -0.575 \\ 0.139 & -3.126 & 0.121 & -0.575 \\ 0.121 & 0.121 & -9.477 & -0.121 \\ -0.575 & -0.575 & -0.121 & 0.139 \end{pmatrix} \begin{matrix} \Delta Z_1 \\ \Delta Z_2 \\ \Delta \mathbf{R} \\ \Delta N \end{matrix} \quad (4)$$

The corresponding eigenvalues and eigenvectors are

$$\begin{pmatrix} \text{ANM}_1 & \text{ANM}_2 & \text{ANM}_3 & \text{ANM}_4 \\ -0.018 & 0.168 & 0.707 & 0.687 \\ -0.018 & 0.168 & -0.707 & 0.687 \\ 1. & 0.016 & -0. & 0.022 \\ 0.01 & -0.971 & -0. & 0.238 \\ -9.48 & 0.34 & -3.27 & -3.18 \end{pmatrix} \begin{matrix} \Delta Z_1 \\ \Delta Z_2 \\ \Delta \mathbf{R} \\ \Delta N \\ \epsilon \end{matrix} \quad (5)$$

where ANM_1 and ANM_2 are mostly changing \mathbf{R} and N with eigenvalues -9.48 [Ha/bohr²] and 0.34 [Ha/ e^2] respectively. ANM_3 (eigenvalue -3.27 [Ha/ e^2]) is purely antisymmetric in changes in Z_1 and Z_2 . ANM_4 (eigenvalue -3.18 [Ha/ e^2]) is remarkable: one can see how in this principle component the increase of the nuclear charge on both atoms requires also an increase in number of electrons (to compensate the change in Z) *and* requires a slight increase of the distance between the atoms. It should be

noted that due the seminegative definite nature of the linear response function, the eigenvalues with respect to the changes in \mathbf{R} , Z are all negative.

6.1 Computational details

For the diatomics and coronene, we used PBE⁵⁷ with uncontracted cc-pVDZ Ne basis function for all atoms in Gaussian09⁵⁸ and HORTON⁵⁹ for all molecular examples.

For benzene, we used CCSD/cc-pvdz in Gaussian09⁵⁸ with the `massage` keyword to modify the nuclear charges. As such, the Carbon cc-pvdz basis functions were also used for Nitrogen and Boron (uncontracting it was computationally very expensive and created instabilities we could not resolve). To calculate the curvature, Benzene was calculated $\pm 0.25 \times$ eigen vector. So for the completely symmetric A1g, this means the carbon nuclear charge $+0.25 \times 1/\sqrt{6} = 6.1020620725$.

For the energy calculations of solids, a 1x1x1 face-centered cubic (fcc) primitive super cell with two atoms was used (no k-point sampling) within the plane-wave basis set code CPMD,⁶⁰ in combination with the PBE⁵⁷ functional, a plane-wave cutoff of 100 Ha, and Goedecker-Teter-Hutter pseudopotentials.^{55,56}

6.2 ANM based estimates

One can expand the potential energy ground state hyper surface of any target system $E_t(\{\mathbf{R}, Z\})$ around a symmetric iso-electronic reference system with identical atomic coordinates and energy E_0 by means of a Taylor expansion in coupling parameter $0 \leq \lambda \leq 1$,

$$E_t(\{\mathbf{R}, Z\}) \approx E_0 + \frac{1}{2} \frac{\partial^2 E_0}{\partial \lambda^2} d\lambda^2 + \text{EHOT} \quad (6)$$

$$E_t(\{\mathbf{R}, Z\}) \approx E_0 + \frac{1}{2} \sum_{IJ} \frac{\partial^2 E_0}{\partial Z_I \partial Z_J} dZ_J dZ_I + \text{EHOT} \quad (7)$$

for $\partial_\lambda Z_J = dZ_J$ and for $d\lambda = 1$, and EHOT corresponding to even higher order terms. Note that instead of nuclear charges $\{Z_I\}$ one can use

pseudopotential parameters $\{\sigma_i\}$ just as well. To simplify this equation, let \mathbf{Q} be the matrix of the eigenvectors of the second order derivative matrix $\mathbf{H}_{IJ} = \frac{\partial^2 E_0}{\partial Z_I \partial Z_J}$ (or Hessian),

$$\mathbf{H}\mathbf{Q} = \mathbf{Q}\mathbf{\Upsilon} \quad (8)$$

where $\mathbf{\Upsilon}$ is a diagonal matrix with the eigenvalues ϵ_m of \mathbf{H} . The eigenvalues ϵ_m are solutions of the alchemical secular equation,

$$\det(\partial_{Z_I, Z_J}^2 E_0 - \delta_{IJ} \epsilon_m) = 0 \quad (9)$$

We now define the alchemical normal mode vector Q_i consisting of the columns of \mathbf{Q} . For a given target molecule, changes in the nuclear charge vector ($d\mathbf{z} = \partial_\lambda \mathbf{z}(\lambda)$), can be expressed in the new basis of the alchemical normal modes as a linear combination:

$$\mathbf{c} = \mathbf{Q}d\mathbf{z} \quad (10)$$

resulting in

$$\sum_{IJ} \frac{\partial E_0}{\partial Z_I \partial Z_J} dZ_I dZ_J = \sum_i \epsilon_i c_i^2 \quad (11)$$

The energy along a chosen alchemical path $\mathbf{z}(\lambda)$ can be expressed as $E(\mathbf{z}(\lambda))$ where the corresponding alchemical derivative at the reference system 0 is

$$\partial_\lambda U_0(\mathbf{z}(\lambda)) = \nabla E_0(\mathbf{z}) \cdot d\mathbf{z} = \sum_I (\partial_{Z_I} E_0) \partial_\lambda Z_I, \quad (12)$$

where $E_0(\mathbf{z}) = E_0(Z_1, \dots, Z_N)$ is a $\mathbf{R}^N \mapsto \mathbf{R}$ function described by $\{Z_I\}$.

Within the orthogonalization transformation, the basis is changed from nuclear charges $\{Z_I\}$ to alchemical normal modes $\{Q_i\}$ where the magnitude in each dimension c_i denotes the amplitude of each normal mode. In other words, the energy is rewritten as

$$E_0(\mathbf{z}) \Rightarrow E_0(\mathbf{c}) = E_0(c_1, \dots, c_N). \quad (13)$$

Notice that \mathbf{c} is a linear function in λ due to Eq. (10) where \mathbf{Q} is independent of λ and \mathbf{z} is linear in λ .

The alchemical derivative within alchemical

normal mode basis is

$$\partial_\lambda E_0(\mathbf{c}(\lambda)) = \sum_i (\partial_{c_i} E_0) \partial_\lambda c_i. \quad (14)$$

Notice that $\partial_\lambda c_i = \sum_J Q_{iJ} \partial_\lambda Z_J$. And the second order derivative is

$$\begin{aligned} \partial_\lambda^2 E_0 &= \partial_\lambda \left(\sum_i \frac{\partial E_0}{\partial c_i} \frac{\partial c_i}{\partial \lambda} \right) \\ &= \sum_i \left(\sum_j \frac{\partial}{\partial c_j} \frac{\partial E_0}{\partial c_i} \frac{\partial c_j}{\partial \lambda} \right) \frac{\partial c_i}{\partial \lambda} \\ &= \sum_{ij} \frac{\partial^2 E_0}{\partial c_i \partial c_j} (\partial_\lambda c_i) (\partial_\lambda c_j). \end{aligned} \quad (15)$$

Notice that $\partial_\lambda^2 c_i = 0$ because $d\mathbf{z}$ is linear in λ . And $\frac{\partial^2 E_0}{\partial c_i \partial c_j} = \delta_{ij} \epsilon_i$ is the diagonal matrix and it is connected to $\frac{\partial^2 E_0}{\partial Z_I \partial Z_J}$ via

$$\left(\frac{\partial^2 E_0}{\partial c_i \partial c_j} \right)_{ij} = \mathbf{Q}^T \left(\frac{\partial^2 E_0}{\partial Z_I \partial Z_J} \right)_{IJ} \mathbf{Q} \quad (16)$$

And the Eq. (11) can be rewritten as

$$\begin{aligned} \sum_{IJ} \frac{\partial E_0}{\partial Z_I \partial Z_J} dZ_I dZ_J &= d\mathbf{z}^T \left(\frac{\partial^2 E_0}{\partial Z_I \partial Z_J} \right)_{IJ} d\mathbf{z} \\ &= d\mathbf{z}^T \mathbf{Q}^T \left(\frac{\partial^2 E_0}{\partial c_i \partial c_j} \right)_{ij} \mathbf{Q} d\mathbf{z} \\ &= \mathbf{c}^T \left(\frac{\partial^2 E_0}{\partial c_i \partial c_j} \right)_{ij} \mathbf{c} \\ &= \sum_i \epsilon_i c_i^2 \end{aligned} \quad (17)$$

6.3 Pseudopotential space

When pseudopotentials (PP) are used, ANM space is spanned by the PP parameters. For fcc primitive cell of two atoms, there are 30 parameters (15 per atoms). The Hessian matrix elements can be approximate by finite difference

$$\begin{aligned} \mathbf{H}_{ij} &= \frac{\partial^2 E}{\partial \sigma_i \partial \sigma_j} \\ &= \frac{\partial}{\partial \sigma_i} \left(\frac{\partial E}{\partial \sigma_j} \right) \\ &\approx \frac{\partial}{\partial \sigma_i} \left(\frac{E(\sigma_i, \sigma_j + \Delta \sigma_j) - E(\sigma_i, \sigma_j)}{\Delta \sigma_j} \right) \\ &\approx \frac{1}{\Delta \sigma_i} \left(\frac{E(\sigma_i + \Delta \sigma_i, \sigma_j + \Delta \sigma_j) - E(\sigma_i + \Delta \sigma_i, \sigma_j)}{\Delta \sigma_j} \right. \\ &\quad \left. - \frac{E(\sigma_i, \sigma_j + \Delta \sigma_j) - E(\sigma_i, \sigma_j)}{\Delta \sigma_j} \right). \end{aligned} \quad (18)$$

That is, there are four finite difference calculations required for each of the matrix elements: $E(\sigma_i + \Delta \sigma_i, \sigma_j + \Delta \sigma_j)$, $E(\sigma_i + \Delta \sigma_i, \sigma_j)$, $E(\sigma_i, \sigma_j + \Delta \sigma_j)$, $E(\sigma_i, \sigma_j)$, where only the first term is unique for each element.

Note that the finite difference formula is different for diagonal terms $\mathbf{H}_{ii} = \frac{E(\sigma_i + \Delta \sigma_i) - 2E(\sigma_i) + E(\sigma_i - \Delta \sigma_i)}{\Delta \sigma_i^2}$. The required finite difference calculations are

- $E(\sigma_i + \Delta \sigma_i, \sigma_j + \Delta \sigma_j)$: $N(N-1)/2$ calculations for $i \neq j$.
- $E(\sigma_i + \Delta \sigma_i)$: N calculations.
- $E(\sigma_i - \Delta \sigma_i)$: N calculations.
- $E(\sigma_i, \sigma_j)$: 1 calculation

which adds up to $\frac{N^2}{2} + \frac{3}{2}N + 1$ calculations, where N is the number of parameters in the system.

7 Acknowledgements

We would like to thank F. A. Faber and G. F. von Rudorff for insightful discussions. OAvL acknowledges support by the Swiss National Science foundation (No. PP00P2_138932, 407540_167186 NFP 75 Big Data, 200021_175747, NCCR MARVEL). Some calculations were performed at sciCORE (<http://scicore.unibas.ch/>) scientific computing core facility at University of Basel.

References

- (1) Hafner, J.; Wolverton, C.; Ceder, G.; Editors, G. Toward computational materials design: The impact of density functional theory on materials research. *MRS Bulletin* **2006**, *31*, 659.
- (2) Jain, A.; Hautier, G.; Moore, C. J.; Ong, S. P.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A high-throughput infrastructure for density functional theory calculations. *Computational Materials Science* **2011**, *50*, 2295–2310.
- (3) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nature Mater* **2013**, *12*, 191–201.
- (4) Ong, S. P.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Bailey, D.; Skinner, D.; Persson, K. A.; Ceder, G. The Materials Project. 2011; <http://materialsproject.org/>, <http://materialsproject.org/>.
- (5) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *Jom* **2013**, *65*, 1501–1509.
- (6) Ramakrishnan, R.; Dral, P.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **2014**, *1*, 140022.
- (7) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific data* **2017**, *4*, 170193.
- (8) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823.
- (9) Mullard, A. The drug-maker’s guide to the galaxy. *Nature News* **2017**, *549*, 445.
- (10) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
- (11) von Lilienfeld, O. A. Quantum Machine Learning in Chemical Compound Space. *Angewandte Chemie International Edition* **2018**, *57*, 4164, <http://dx.doi.org/10.1002/anie.201709686>.
- (12) Paliana, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Scientific reports* **2013**, *3*, 2810.
- (13) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **2014**, *89*, 094104.
- (14) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **2014**, *89*, 205118.
- (15) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals. *Phys. Rev. Lett.* **2016**, *117*, 135502.
- (16) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.

- (17) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics* **2018**, *148*, 241717.
- (18) von Lilienfeld, O. A. First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties. *International Journal of Quantum Chemistry* **2013**, *113*, 1676–1689.
- (19) Hückel, E. Quantentheoretische beiträge zum benzolproblem. *Zeitschrift für Physik* **1931**, *70*, 204–286.
- (20) Wheland, G. W.; Pauling, L. A quantum mechanical discussion of orientation of substituents in aromatic molecules. *Journal of the American Chemical Society* **1935**, *57*, 2086–2095.
- (21) Coulson, C. A.; Longuet-Higgins, H. C. The electronic structure of conjugated systems I. General theory. *Proc. R. Soc. Lond. A* **1947**, *191*, 39–60.
- (22) E. B. Wilson, Jr., Four Dimensional Electron Density Function. *J. Chem. Phys.* **1962**, *36*, 2232.
- (23) Politzer, P.; Parr, R. G. Some new energy formulas for atoms and molecules. *J. Chem. Phys.* **1974**, *61*, 4258.
- (24) Mezey, P. G. A simple relation between nuclear charges and potential surfaces. *J. Am. Chem. Soc.* **1985**, *107*, 3100.
- (25) Marzari, N.; de Gironcoli, S.; Baroni, S. Structure and Phase Stability of $\text{Ga}_x\text{In}_{1-x}\text{P}$ solid solutions from computational Alchemy. *Phys. Rev. Lett.* **1994**, *72*, 4001.
- (26) Sablon, N.; de Proft, F.; Ayers, P. W.; Geerlings, P. Computing Second-Order Functional Derivatives with Respect to the External Potential. *J. Chem. Theory Comput.* **2000**, *6*, 3671.
- (27) von Lilienfeld, O. A.; Lins, R.; Rothlisberger, U. Variational particle number approach for rational compound design. *Phys. Rev. Lett.* **2005**, *95*, 153002.
- (28) Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. Designing molecules by optimizing potentials. *J. Am. Chem. Soc.* **2006**, *128*, 3228.
- (29) Beste, A.; Harrison, R. J.; Yanai, T. Direct computation of general chemical energy differences: Application to ionization potentials, excitation, and bond energies. *J. Phys. Chem.* **2006**, *125*, 074101.
- (30) von Lilienfeld, O. A. Accurate *ab initio* gradients in chemical compound space. *J. Chem. Phys.* **2009**, *131*, 164102.
- (31) Sheppard, D.; Henkelman, G.; von Lilienfeld, O. A. Alchemical derivatives of reaction energetics. *J. Chem. Phys.* **2010**, *133*, 084104.
- (32) Lesiuk, M.; Balawender, R.; Zachara, J. Higher order alchemical derivatives from coupled perturbed self-consistent field theory. *J. Chem. Phys.* **2012**, *136*, 034104.
- (33) Balawender, R.; Welearegay, M. A.; Lesiuk, M.; De Proft, F.; Geerlings, P. Exploring Chemical Space with the Alchemical Derivatives. *J. Chem. Theory Comput.* **2013**, *9*, 5327–5340.
- (34) Chang, K. Y. S.; von Lilienfeld, O. A. Quantum Mechanical Treatment of Variable Molecular Composition: From “Alchemical” Changes of State Functions to Rational Compound Design. *CHIMIA* **2014**, *1*.
- (35) Chang, K. Y. S.; Fias, S.; Ramakrishnan, R.; von Lilienfeld, O. A. Fast and accurate predictions of covalent bonds in chemical space. *J. Chem. Phys.* **2016**, *144*, 174110.
- (36) Solovyeva, A.; von Lilienfeld, O. A. Alchemical screening of ionic crystals. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31078–31091.

- (37) to Baben, M.; Achenbach, J. O.; von Lilienfeld, O. A. Guiding ab initio calculations by alchemical derivatives. *J. Chem. Phys.* **2016**, *144*, 104103.
- (38) Al-Hamdani, Y. S.; Michaelides, A.; von Lilienfeld, O. A. Exploring water adsorption on isoelectronically doped graphene using alchemical derivatives. *J. Chem. Phys.* **2017**, *147*, 164113, <http://arxiv.org/abs/1703.10083>.
- (39) Saravanan, K.; Kitchin, J. R.; von Lilienfeld, O. A.; Keith, J. A. Alchemical Predictions for Computational Catalysis: Potential and Limitations. *The Journal of Physical Chemistry Letters* **2017**, *8*, 5002–5007.
- (40) Fias, S.; Heidar-Zadeh, F.; Geerlings, P.; Ayers, P. W. Chemical transferability of functional groups follows from the near-sightedness of electronic matter. *Proceedings of the National Academy of Sciences* **2017**, *114*, 11633–11638.
- (41) Balawender, R.; Lesiuk, M.; De Proft, F.; Geerlings, P. Exploring Chemical Space with Alchemical Derivatives: BN-Simultaneous Substitution Patterns in C60. *Journal of chemical theory and computation* **2018**,
- (42) Tuckerman, M. E. *Statistical mechanics: Theory and molecular simulation*; Oxford University Press, 2010.
- (43) von Lilienfeld, O. A.; Tuckerman, M. E. Molecular grand-canonical ensemble density functional theory and exploration of chemical space. *J. Chem. Phys.* **2006**, *125*, 154104.
- (44) Marcon, V.; von Lilienfeld, O. A.; Andrienko, D. Tuning electronic eigenvalues of benzene via doping. *J. Chem. Phys.* **2007**, *127*, 064305.
- (45) von Lilienfeld, O. A.; Tuckerman, M. E. Alchemical variation of intermolecular energies according to molecular grand-canonical ensemble density functional theory. *J. Chem. Theory Comput.* **2007**, *3*, 1083.
- (46) Janak, J. F. Proof that $\partial E/\partial n_i = \epsilon_i$ in density-functional theory. *Phys. Rev. B* **1978**, *18*, 7165.
- (47) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz, J. L. *Phys. Rev. Lett.* **1982**, *49*, 1691.
- (48) Perdew, J. P.; Levy, M. *Phys. Rev. Lett.* **1983**, *51*, 1884.
- (49) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Discontinuous nature of the exchange-correlation functional in strongly correlated systems. *Phys. Rev. Lett.* **2009**, *102*, 066403.
- (50) Parr, R. G.; Yang, W. *Density functional theory of atoms and molecules*; Oxford Science Publications, 1989.
- (51) Baekelandt, B. G. The nuclear Fukui function and Berlin’s binding function in density functional theory. *The Journal of chemical physics* **1996**, *105*, 4664–4667.
- (52) Andrienko, D.; Kirkpatrick, J.; Marcon, V.; Nelson, J.; Kremer, K. Structure-charge mobility relation for hexabenzocoronene derivatives. *Phys. Stat. Sol. B* **2008**, *245*, 830.
- (53) Feng, X.; Marcon, V.; Pisula, W.; Hansen, M. R.; Kirkpatrick, J.; Grozema, F.; Andrienko, D.; Kremer, K.; Mullen, K. Rational design of the shape and periphery of discotics: a synthetic way towards high charge carrier mobilities. *Nature Materials* **2009**, *8*, 421.
- (54) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

- (55) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703.
- (56) Krack, M. *Theor. Chim. Acta* **2005**, *114*, 145–152.
- (57) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (58) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Á.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT 2009.
- (59) Verstraelen, T.; Tecmer, P.; Heidarzadeh, F.; González-Espinoza, C. E.; Chan, M.; Kim, T. D.; Boguslawski, K.; Fias, S.; Vandenbrande, S.; Berrocal, D.; ; Ayers, P. W. HORTON. **2017**, <http://theochem.github.com/horton/>.
- (60) J. Hutter et al., Computer code CPMD,