

LETTER • OPEN ACCESS

Wasserstein metric for improved quantum machine learning with adjacency matrix representations

To cite this article: Onur Çaylak *et al* 2020 *Mach. Learn.: Sci. Technol.* 1 03LT01

View the [article online](#) for updates and enhancements.

You may also like

- [Hardening quantum machine learning against adversaries](#)
Nathan Wiebe and Ram Shankar Siva Kumar
- [Hybrid functionals and GW approximation in the FLAPW method](#)
Christoph Friedrich, Markus Betzinger, Martin Schlipf *et al.*
- [Machine learning of molecular electronic properties in chemical compound space](#)
Grégoire Montavon, Matthias Rupp, Vivekanand Gobre *et al.*



EDINBURGH
INSTRUMENTS

WORLD LEADING
MOLECULAR
SPECTROSCOPY SOLUTIONS



edinst.com



LETTER

OPEN ACCESS

RECEIVED
30 January 2020REVISED
15 June 2020ACCEPTED FOR PUBLICATION
26 June 2020PUBLISHED
11 August 2020

Original Content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



Wasserstein metric for improved quantum machine learning with adjacency matrix representations

Onur Çaylak^{1,2,3}, O. Anatole von Lilienfeld^{4,3} and Björn Baumeier^{1,2} ¹ Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands² Institute for Complex Molecular Systems, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands³ Institute for Pure and Applied Mathematics, University of California Los Angeles, 460 Portola Plaza, Los Angeles, California 90095, United States of America⁴ Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), University of Basel, Department of Chemistry, 4056 Basel, SwitzerlandE-mail: anatole.vonlilienfeld@unibas.ch and b.baumeier@tue.nl**Keywords:** Quantum Machine Learning, Adjacency Matrix Representation, Kernel Ridge Regression, Atomization Energies, Wasserstein metric

Abstract

We study the Wasserstein metric to measure distances between molecules represented by the atom index dependent adjacency ‘Coulomb’ matrix, used in kernel ridge regression based supervised learning. Resulting machine learning models of quantum properties, a.k.a. quantum machine learning models exhibit improved training efficiency and result in smoother predictions of energies related to molecular distortions. We first illustrate smoothness for the continuous extraction of an atom from some organic molecule. Learning curves, quantifying the decay of the atomization energy’s prediction error as a function of training set size, have been obtained for tens of thousands of organic molecules drawn from the QM9 data set. In comparison to conventionally used metrics (L_1 and L_2 norm), our numerical results indicate systematic improvement in terms of learning curve off-set for random as well as sorted (by norms of row) atom indexing in Coulomb matrices. Our findings suggest that this metric corresponds to a favorable similarity measure which introduces index-invariance in any kernel based model relying on adjacency matrix representations.

1. Introduction

The application of machine learning (ML) to atomistic simulation has been gaining traction over recent years [1–9]. Kernel ridge regression (KRR) models of quantum properties (Q) applicable throughout chemical compound space (CCS) was established in 2012 [10], and has been growing ever since [11–15]. See references [16, 17] for more details and references about the methodology known as Quantum Machine Learning (QML). By now, QML has become a viable and popular approach for generating surrogate property models enabling rapid estimates of relevant molecular and materials properties, holding great promise for computational materials and molecular design [18, 19], as recently exemplified for the discovery of nearly ninety stable crystal candidates in the Elpasolite structure [20].

When setting up standard KRR based QML models of some quantum property P (aka ‘label’) [13], three fundamental choices must be made, (i) the representation \mathbf{x} (aka ‘feature’), (ii) the kernel function (k), and (iii) the metric ($\text{dist}(\cdot, \cdot)$), such that

$$P(\mathbf{x}) = \sum_i^N \beta_i k(\text{dist}(\mathbf{x}, \mathbf{x}_i)), \quad (1)$$

where N and $\{\beta_i\}$ correspond to number and regression coefficients of training instances, respectively. The representation of a chemical system is known to play an important role. For example, when using incomplete

representations (or non-unique), proof was given that QML models can produce absurd results [21]. While the details of the representation (other than uniqueness) are less crucial for artificial neural networks, the specific definition of how a chemical system is being specified is known to dramatically affect the learning efficiency of KRR based QML models. Namely, encoding the right physics, such as translational or atom-index invariance, in the representation results in systematic reduction of quantum data needs for achieving the same pre-defined predictive accuracy [22]. This is of particular interest since QML models are typically trained within scarce data regimes due to (a) the immense computational (or experimental) cost for generating labels and (b) the tremendous scale of CCS [23, 24].

Due to their obvious impact on QML model performance, it is not surprising that substantial efforts have been made to improve conventional representations. For example, using atomization energies of organic molecules stored in the QM9 dataset [25], various benchmark results have been obtained including as representations the Coulomb matrix and BOB (2015) [26], BAML (2016) [22], HDAD (2017) [27], constant-size-descriptors (2018) [28], SLATM [29] (2017), SOAP (2017) [16], FCHL (2018) [30], MBD (2018) [31], and wavelets (2018) [32]). See reference [9] for a joint graphical illustration of learning curves coming from these, as well as artificial neural network based, models. Apart from the representation, the choice of kernel function is also known to affect the performance of the QML model, as shown in references [13, 16, 33–35].

Within this paper, we focus on the aforementioned third choice (iii): The metric. More specifically, previous studies have predominantly relied on Euclidean or Manhattan norms as a metric. This choice can be questioned when it comes to atom index dependence of adjacency matrices, such as the Coulomb matrix (CM) [10], possibly resulting in discontinuities in the surrogate model due to displacement (or alchemical change [24]) of the nucleus. In this paper, we discuss how such spurious artefacts are resolved by using a more sophisticated, distribution based measure: The Wasserstein metric [36].

2. Method

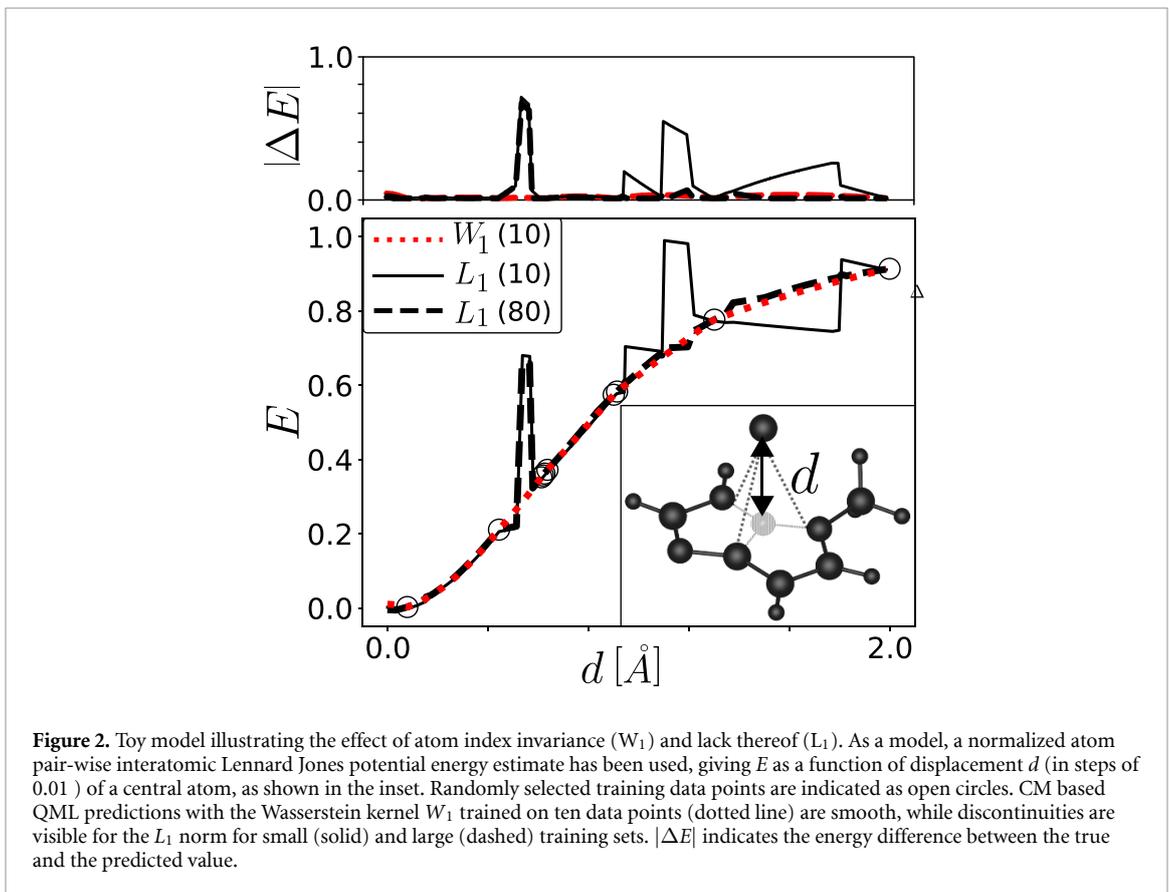
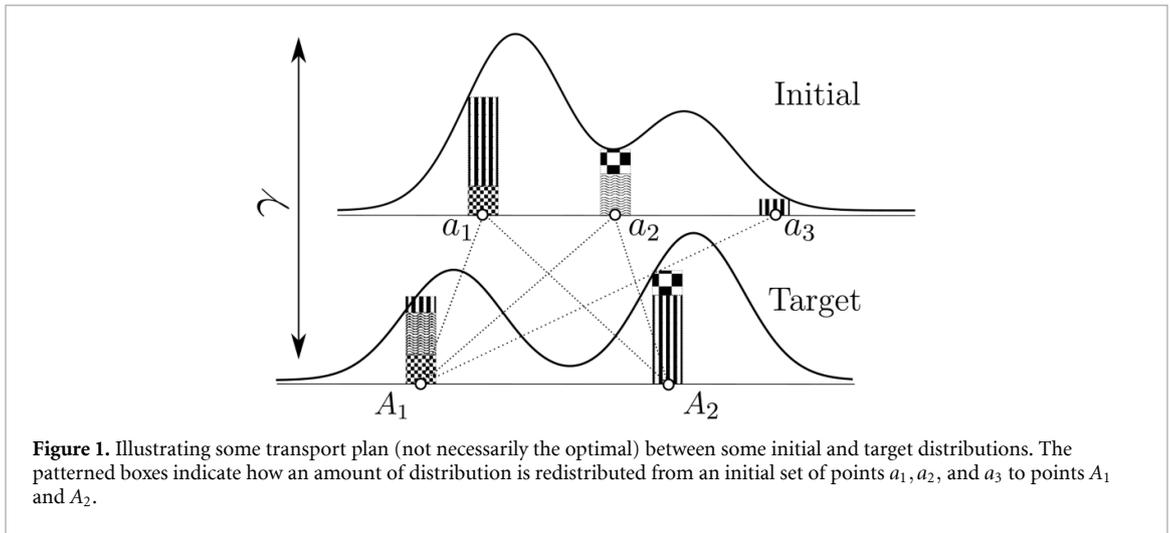
The CM is an adjacency matrix with diagonal and off-diagonal terms corresponding to approximate free atom and nuclear repulsion contributions to the total potential energy of a molecule, respectively [10]. Its adaptation to crystal representations was published subsequently [37]. Its creation was motivated by the fact that it is unique for fixed molecular charges up to permutation of atoms, and that first-principles calculations also require only nuclear coordinates \mathbf{R}_I and nuclear charges Z_I as input.

$$C_{IJ} = \begin{cases} 0.5Z_I^{2.4} & I = J \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & I \neq J \end{cases} \quad (2)$$

As the CM is invariant to 3D translations and rotations of a molecule, it intrinsically ensures that the molecule's potential energy is constant under those transformations. Among the early problems identified for the general application of the Coulomb matrix is the index dependence. Sorting the Coulomb matrix such that $\|C_m\| \geq \|C_{m+1}\|$ for all m , where C_m is the m th row, renders the CM bijective up to rotation and translation. While appealing for its simplicity and still in use for various applications [38, 39], the use of the sorted CM must be cautioned when applied to situations in which smooth geometrical (or alchemical) changes are under consideration. Such applications include training throughout CCS and subsequent prediction of energies in molecular dynamics trajectories, or geometry relaxations energies, when sorting can lead to a swapping of indices in the vectorized forms of the CM, and sudden atomic index reassignments between test and query system. This *indexing problem* manifests itself in discontinuities in predictions and/or the need for a large number of data points for training the respective models.

Here, rather than attempting to resolve the indexing problem through ever more sophisticated representations, for example using atom centered symmetry functions [40], SOAP [41], HDAD [27], SLATM [29], MBD [31], or FCHL [30], we investigate if this issue can also be resolved by using a different metric, capable of alleviating the sudden re-assignment occurring within L_1 or L_2 norms. In particular, the Wasserstein norm of order 1, which is denoted by W_1 , is a natural way to compare two probability distributions p and q [36]. This norm is widely used in various fields, like machine learning, image processing, and signal processing [42–44].

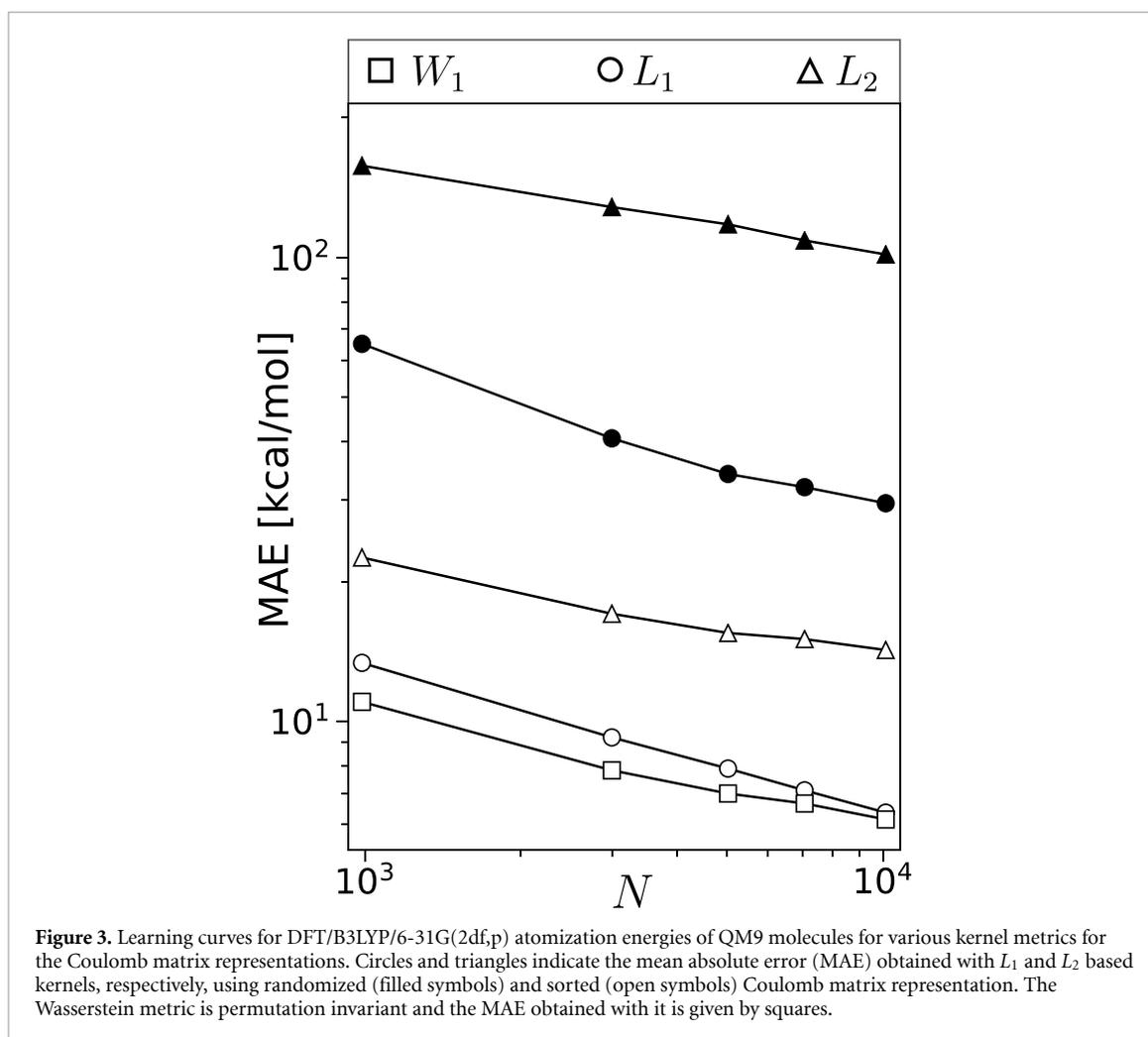
As illustrated in figure 1, the Wasserstein metric is the minimal amount of work needed to transform one distribution into another; work being defined as the amount of distribution times the distance it has to be transported. There are many different ways of transporting an amount of distribution from a region x of p into a region y of q . The set of all possible transport plans to move p into q is denoted by Γ . Hence, computing the distance between two distributions can be formulated as an optimization problem where the aim is to find that transport plan $\gamma \in \Gamma$ such that the total amount of work is minimal. The Wasserstein [36]



metric is expressed as

$$\begin{aligned}
 W_1(p, q) &= \inf_{\gamma \in \Gamma} \int_{\mathbb{R}^d \times \mathbb{R}^d} d\gamma(x, y) |x - y| \\
 &= \int_{\mathbb{R}} dt |P^{-1}(t) - Q^{-1}(t)|,
 \end{aligned} \tag{3}$$

where the right-hand side equation was shown to hold in reference [45] with P and Q being the cumulative distribution functions of p and q . The vectorized two-dimensional Coulomb matrix representation of a molecule can be used in equation (3). In practice, the evaluation is done by using empirical cumulative distribution functions (cdf). The integral in equation(3) becomes the absolute value of the difference of two distinct cdf's. So, in worst case scenario it behaves like $\mathcal{O}(2n \log(n))$. From now on, we set the L_1 and L_2 based kernel functions k to be the well known Laplacian and Gaussian kernel, respectively. Additionally, we



define the Wasserstein-based kernel as

$$k(\mathbf{x}, \mathbf{y}) = e^{-\alpha W_1(\mathbf{x}, \mathbf{y})}. \quad (4)$$

Here, we note that other kernel functions or representations could be used in combination with the Wasserstein metric just as well.

All QML models of atomization energies of QM9 molecules [25] were trained using KRR with 5-fold cross-validation for hyperparameter optimization, and tested on 2000 out-of-sample molecules.

3. Results and discussion

First, we illustrate the issue of smoothness by subjecting an innocent organic molecule to drastic distortions. More specifically, and as shown in figure 2, consider the energy E as a function of continuous displacement d (in steps of 0.01) of some central carbon atom along an axis orthogonal to the molecular plane. The molecule used in this example is drawn randomly from the QM9 dataset¹. Corresponding Lennard-Jones potential is calculated using the LJ formula where all parameters are set to 1 independent of atom type. Furthermore, we have normalized the curves with respect to the highest energy of obtained results. The energies are smooth, while CM based QML model predictions (after training on 10 instances drawn at random) using L_1 are discontinuous. Even after increasing the training set for the L_1 model to 80 instances, the discontinuity at $d \approx 0.5$ is retained, indicating that lack of learning. It is non trivial to figure out what the exact cause is for the initial discontinuity, mainly due to the highly dependent variables in the CM representation. The carbon atom that is being extracted is bonded to two nitrogen atoms and one carbon atom. However, there is another carbon–nitrogen bond in the molecule with initially a longer bond length that is not bonded to the extracted carbon atom. The extraction of the carbon atom stretches the bonds and around $d \approx 0.5$ the order

¹dsgdb9nsd_030000.xyz.

swaps, which could be the reason for the first discontinuity in the curve. As pointed out above, the main benefit of equation (4) is that W_1 is invariant under permutations of row- and column-indices of the adjacency matrix. When applied to the model system from figure 2, which showed the typical indexing problem with the Laplacian kernel and L_1 -norm, the W_1 metric yields smooth energies as a function of displacement. Two aspects are noteworthy: First, it is apparent that the predictions are smooth (differentiable) across the full range of displacements and do not exhibit any discontinuities that affected the standard kernels in combination with the sorted Coulomb matrix. Second, a training data set including only 10 reference points is sufficient to give accurate results. These numerical results clearly indicate that the use of the Wasserstein metric in the kernel cures the indexing problem and alleviates the associated low prediction quality and data efficiency.

To further investigate the performance of Wasserstein based kernels as in equation (4) in QML models, we have turned to the classic benchmark of atomization energies of organic molecules in the QM9 data set [25]. After training on up to 10k molecules, randomly sampled from the entire QM9 data set, learning curves are presented in figure 3. Results for L_2 and L_1 based kernels are in line with observations made in reference [13]: For sorted as well as for randomly indexed CMs the L_1 based QML model exhibits lower off-sets than the L_2 based model. Not surprisingly, use of sorted CMs also results in smaller off-sets than for random CMs. However, the W_1 based metric results in the same learning curve after being applied to random as well as to sorted CMs. Its overall learning curve off-set and slope indicates same (even slightly better) performance as the L_1 norm applied to the sorted CM, reaching ~ 6 kcal mol⁻¹ after training on 10k instances.

4. Conclusions

Considering the findings for the continuous atomic displacement as well as the QM9 molecules, the Wasserstein metric enables the generation of QML models which achieve (a) data-efficient learning and (b) smooth target function estimates. While all our numerical evidence has relied on the CM, we emphasize that the observed solution of the indexing problem and the simultaneous improvement of the predictions by using the Wasserstein kernel is not inherently specific to the CM representation. In fact, the Wasserstein metric can readily be tested with other QML models relying on any graph-based representation. This could be particularly relevant in the context of recent work on learning force-fields or electronic properties, relying on inverse distances rather than the CM representation. [47–50]

To summarize, we have presented the Wasserstein metric as an index-invariant way to measure distances between molecular graph-based representations. Our numerical findings indicate that the resulting QML models combine smoothness with data efficiency in learning. Future work will explore the various possible combinations of kernel functions, Wasserstein metric, and representations other than the CM. All methods used in this work are now also available in the QMLCODE package. [51]

Acknowledgments

O.A.v.L. acknowledges funding from the Swiss National Science foundation (No. 407540_167186 NFP 75 Big Data) and from the European Research Council (ERC-CoG grant QML). This work was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. B. B. acknowledges support by the Innovational Research Incentives Scheme Vidi of the Netherlands Organisation for Scientific Research (NWO) with project number 723.016.002. Partial funding is also provided by NWO and the Netherlands eScience Center for funding through project number 027.017.G15, within the Joint CSER and eScience program for Energy Research (JCER 2017).

Data availability statement

No new data were created or analysed in this study.

ORCID iD

Björn Baumeier  <https://orcid.org/0000-0002-6077-0467>

References

- [1] Behler J 2015 Constructing high-dimensional neural network potentials: a tutorial review *Int. J. Quantum Chem.* **115** 1032–50
- [2] Rupp M 2015 Special issue on machine learning and quantum mechanics *Int. J. Quantum Chem.* **115** 1003–4
- [3] Behler J 2016 Perspective: Machine learning potentials for atomistic simulations *J. Chem. Phys.* **145** 170901

- [4] Rupp M, von Lilienfeld O A and Burke K 2018 Guest editorial: special topic on data-enabled theoretical chemistry *J. Chem. Phys.* **148** 241401
- [5] Noé F, Tkatchenko A, Müller K-R and Clementi C 2020 Machine learning for molecular simulation *arXiv preprint Annu. Rev. Phys. Chem.* **71** 361–90
- [6] Musil F and Ceriotti M 2019 Machine learning at the atomic scale *CHIMIA Int. J. Chem.* **73** 972–82
- [7] Ceriotti M 2019 Unsupervised machine learning in atomistic simulations, between predictions and understanding *J. Chem. Phys.* **150** 150901
- [8] Ceriotti M, Willatt M J and Csányi G 2018 Machine learning of atomic-scale properties based on physical principles *Handbook of Materials Modeling* eds W Andreoni *et al* (Berlin: Springer) (https://doi.org/10.1007/978-3-319-42913-7_68-1)
- [9] von Lilienfeld O A, Tkatchenko A and Müller K-R 2020 Exploring chemical compound space with quantum-based machine learning *Nat. Rev. Chem.* **4** 347–58
- [10] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
- [11] Montavon G, Hansen K, Fazli S, Rupp M, Biegler F, Ziehe A, Tkatchenko A, von Lilienfeld O A and Müller K-R 2012 Learning invariant representations of molecules for atomization energy prediction *NIPS proceedings* **25** 449
- [12] Montavon G, Rupp M, Gobre V, Vazquez-Mayagoitia A, Hansen K, Tkatchenko A, Müller K R and von Lilienfeld O A 2013 Machine learning of molecular electronic properties in chemical compound space *New J. Phys.* **15** 9
- [13] Hansen K, Montavon G, Biegler F, Fazli S, Rupp M, Scheffler M, von Lilienfeld O A, Tkatchenko A and Müller K R 8 2013 Assessment and validation of machine learning methods for predicting molecular atomization energies *J. Chem. Theory Comput.* **9** 3404–19
- [14] Pilia G, Wang C, Jiang X, Rajasekaran S and Ramprasad R 2013 Accelerating materials property predictions using machine learning *Sci. Rep.* **3** 2810
- [15] De S, Bartók A P, Csányi G and Ceriotti M 2016 Comparing molecules and solids across structural and alchemical space *Phys. Chem. Chem. Phys.* **18** 13754–69
- [16] Bartók A P, De S, Poelking C, Bernstein N, Kermode J R, Csányi G and Ceriotti M 2017 Machine learning unifies the modeling of materials and molecules *Sci. Adv.* **3** 12
- [17] von Lilienfeld O A 2018 Quantum machine learning in chemical compound space *Angew. Chem. Int. Ed.* **57** 4164
- [18] von Lilienfeld O A 2014 *Towards the Computational Design of Compounds From First Principles, vol. IX of Mathematical Physics Studies* (Berlin: Springer)
- [19] Marzari N 2016 Materials modelling: The frontiers and the challenges *Nat. Mater.* **15** 381
- [20] Faber F A, Lindmaa A, von Lilienfeld O A and Armiento R 2016 Machine learning energies of 2 million elpasolite (abC_2D_6) crystals *Phys. Rev. Lett.* **117** 135502
- [21] von Lilienfeld O A, Ramakrishnan R, Rupp M and Knoll A 2015 Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties *Int. J. Quant. Chem.* **115** 1084 <http://arxiv.org/abs/1307.2918>
- [22] Huang B and von Lilienfeld O A 2016 Communication: Understanding molecular representations in machine learning: the role of uniqueness and target similarity *J. Chem. Phys.* **145** 161102
- [23] Kirkpatrick P and Ellis C 2004 Chemical space *Nature* **432** 823
- [24] von Lilienfeld O A 2013 First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties *Int. J. Quant. Chem.* **113** 1676–89
- [25] Ramakrishnan R, Dral P, Rupp M and von Lilienfeld O A 2014 Quantum chemistry structures and properties of 134 kilo molecules *Scientific Data* **1** 140022
- [26] Hansen K, Biegler F, von Lilienfeld O A, Müller K-R and Tkatchenko A 2015 Interaction potentials in molecules and non-local information in chemical space *J. Phys. Chem. Lett.* **6** 2326
- [27] Faber F A *et al* 2017 Prediction errors of molecular machine learning models lower than hybrid DFT error *J. Chem. Theory Comput.* **13** 5255–64
- [28] Collins C R, Gordon G J, von Lilienfeld O A and Yaron D J 2018 Constant size descriptors for accurate machine learning models of molecular properties *J. Chem. Phys.* **148** 241718
- [29] Huang B and von Lilienfeld O A 2017 The “DNA” of chemistry: Scalable quantum machine learning with “amons *Arxiv Preprint Arxiv:1707.04146* submitted
- [30] Faber F A, Christensen A S, Huang B and von Lilienfeld O A 2018 Alchemical and structural distribution based representation for universal quantum machine learning *J. Chem. Phys.* **148** 241717
- [31] Pronobis W, Tkatchenko A and Müller K-R 2018 Many-body descriptors for predicting molecular properties with machine learning: analysis of pairwise and three-body interactions in molecules *J. Chem. Theory Comput.* **14** 2991–3003
- [32] Eickenberg M, Exarchakis G, Hirn M, Mallat S and Thiry L 2018 Solid harmonic wavelet scattering for predictions of molecule properties *J. Chem. Phys.* **148** 241732
- [33] Ho T and Rabitz H 1996 A general method for constructing multidimensional molecular potential energy surfaces from abinitio calculations *J. Chem. Phys.* **104** 2584–97
- [34] Soloviov M and Meuwly M 2015 Reproducing kernel potential energy surfaces in biomolecular simulations: nitric oxide binding to myoglobin *J. Chem. Phys.* **143** 09B607_1
- [35] Unke O T and Meuwly M 2017 Toolkit for the construction of reproducing kernel-based representations of data: application to multidimensional potential energy surfaces *J. Chem. Inform. Model.* **57** 1923–31
- [36] Vaserstein L N 1969 Markov processes over denumerable products of spaces, describing large systems of automata *Probl. Pereda. Inf.* **5** 64–72
- [37] Faber F, Lindmaa A, von Lilienfeld O A and Armiento R 2015 Crystal structure representations for machine learning models of formation energies *Int. J. Quantum Chem.* **115** 1094 <http://arxiv.org/abs/1503.07406>
- [38] Çaylak O, Yaman A and Baumeier B 3 2019 Evolutionary approach to constructing a deep feedforward neural network for prediction of electronic coupling elements in molecular materials *J. Chem. Theory Comput.* **15** 1777–84
- [39] Ghosh K, Stuke A, Todorović M, Jørgensen P B, Schmidt M N, Vehtari A and Rinke P 2019 Deep learning spectroscopy: neural networks for molecular excitation spectra *Adv. Sci.* **6** 1801367
- [40] Behler J 2011 Atom-centered symmetry functions for constructing high-dimensional neural networks potentials *J. Chem. Phys.* **134** 074106
- [41] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev. B* **87** 184115

- [42] Frogner C, Zhang C, Mobahi H, Araya-Polo M and Poggio T A 2015 Learning with a Wasserstein loss *Advances in Neural Information Processing Systems 28* eds C Cortes *et al* (Red Hook, NY: Curran Associates, Inc.) pp 2053–61
- [43] Montavon G, Müller K-R and Cuturi M 2015 Wasserstein training of Boltzmann machines *Advances in Neural Information Processing Systems 29* eds D D Lee *et al* (Red Hook, NY: Curran Associates, Inc.) pp 3718–26
- [44] Rabin J, Ferradans S and Papadakis N 2014 Adaptive color transfer with relaxed optimal transport in *2014 IEEE Int. Conf. on Image Processing (ICIP)* pp 4852–6 (New York: IEEE)
- [45] Ramdas A, Trillos N G and Cuturi M 2017 On Wasserstein two-sample testing and related families of nonparametric tests *Entropy* **19** 47
- [46] dsfdb9nsd_030000.xyz.
- [47] Chmiela S, Sauceda H E, Müller K-R and Tkatchenko A 2018 Towards exact molecular dynamics simulations with machine-learned force fields *Nat. Commun.* **9** 3887
- [48] Chmiela S, Sauceda H E, Poltavsky I, Müller K-R and Tkatchenko A 2019 sGDML: Constructing accurate and data efficient molecular force fields using machine learning *Comput. Phys. Commun.* **240** 38–45
- [49] Westermayr J, Faber F A, Christensen A S, von Lilienfeld O A and Marquetand P 2020 Neural networks and kernel ridge regression for excited states dynamics of CH_2NH_2^+ : From single-state to multi-state representations and multi-property machine learning models *Mach. Learn.: Sci. Technol.*, **1** 025009
- [50] Westermayr J, Gastegger M, Menger M F, Mai S, González L and Marquetand P 2019 Machine learning enables long time scale molecular photodynamics simulations *Chem. Sci.* **10** 8100–7
- [51] Christensen A S, Faber F A, Huang B, Bratholm L A, Tkatchenko A, Müller K-R, and von Lilienfeld O A 2017 QML: a python toolkit for quantum machine learning www.qmlcode.org