AABBA: Atom–Atom Bond–Bond Bond–Atom Graph Kernel for Machine Learning on Molecules and Materials

Lucía Morán-González,[†,‡] Jørn Eirik Betten,[‡] Hannes Kneiding,[‡] David Balcells[‡,*]

[†]*Institute of Chemical Research of Catalonia (ICIQ-CERCA), The Barcelona Institute of Science and Technology, Avgda. Països Catalans, 16, 43007 Tarragona, Catalonia, Spain;* [‡]*Hylleraas Centre for Quantum Molecular Sciences, Department of Chemistry, University of Oslo, P.O. Box 1033, Blindern, 0315 Oslo, Norway*

E-mail: david.balcells@kjemi.uio.no

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Abstract

Graphs are one of the most natural and powerful representations available for molecules; natural because they have an intuitive correspondence to skeletal formulas, the language used by chemists worldwide, and powerful, because they are highly expressive both globally (molecular topology) and locally (atomic properties). Graph kernels are used to transform molecular graphs into fixed-length vectors, which can be used as fingerprints in machine learning (ML) models. To date, kernels have mostly focused on the atomic nodes of the graph. In this work, we developed an extended graph kernel computing atom–atom, bond–bond, and bond–atom (AABBA) autocorrelations. The resulting AABBA representations were evaluated with a transition metal complex benchmark, motivated by the higher complexity of these compounds relative to organic molecules. In particular, we tested different flavors of the AABBA kernel in the prediction of the energy barriers and bond distances of the Vaska's complex dataset (Friederich et al., *Chem. Sci.*, **2020**, *11*, 4584). For a variety of ML models, including neural networks, gradient boosting machines, and Gaussian processes, we showed that AABBA outperforms the baseline including only atom–atom autocorrelations. Dimensionality reduction studies also showed that the bond–bond and bond–atom autocorrelations yield many of the most relevant features. We believe that the AABBA graph kernel can accelerate the discovery of chemical compounds and inspire novel molecular representations in which both atomic and bond properties play an important role.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Introduction

Machine learning (ML) is accelerating the fields of catalysis,[1–6] materials science,[7–12] and drug discovery.[13–18] This acceleration is particularly important in the current context defined by the climate and health crises. The further development of deep,[19–21] Bayesian,[22–24] and ensemble[25–28] ML methods is crucial for achieving higher levels of accuracy, generalization, and explainability. Nonetheless, research in the other key components of the ML pipeline, namely data,[29–31] and representations,[32–34] is also crucial, especially in the field of transition metal complex (TMC) chemistry,[35–37] which remains incipient relative to organic chemistry. TMC datasets are smaller and more scarce, and the associated representations can fall short of capturing the intrinsic complexity of these compounds.

The representations commonly used for organic molecules are in general insufficient for TMCs. For example, SMILES[38] and other popular line notations cannot account for the complicated bonding patterns found around metal centers. More advanced notations like SELFIES[39] should in principle overcome this issue but their extension toward TMCs remains unexplored. Graph-based representations are a powerful alternative, due to the higher expressivity of their topology and the possibility of attributing both the nodes (atoms) and edges (bonds) with electronic and geometric data.[40,41] In many ML methods, the graph cannot be directly fed into the model, requiring a function, *i.e.* a graph kernel, which either measures the similarity between graph pairs[42] or, within the focus of this work, transforms the graphs into fixed-length vectors that can be regarded as molecular fingerprints (Figure 1).[43]

The Moreau-Broto autocorrelation[44] is a popular kernel consisting of an algorithm that "walks" over the molecular graph computing atomic property products (Figure 1). The algorithm iterates over the whole graph until an arbitrary depth, which is the maximum number of bonds allowed in the shortest path connecting two autocorrelated nodes. For each property, the products are added to yield the components of the final vector, which has a fixed length defined by the number of properties and the depth of the representation. Ku-

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

lik and co-workers adapted this approach for its application to TMCs by introducing *origin* and *scope* as autocorrelation parameters.[45] The *origin* defines a reference node for which the property products are computed (*e.g.* the metal center), whereas the *scope* delimits the autocorrelation to subgraph sets (*e.g.* axial and equatorial ligands). Further, property products were extended with additional arithmetic operations (*e.g.* subtraction). This implementation is computationally inexpensive and it has proven its efficiency in challenging ML tasks, including the multiobjective optimization of TMCs,[46] though it focuses only on the graph nodes, limiting the extraction of electronic structure data and excluding geometric information.

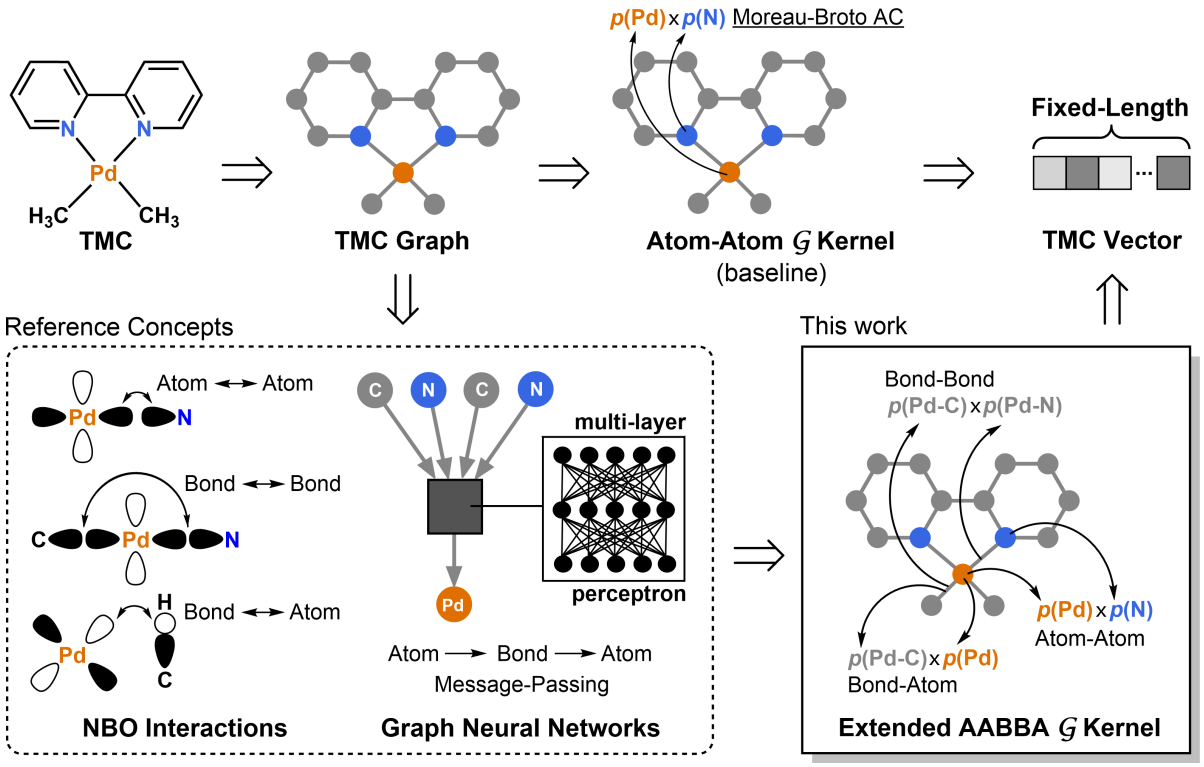

Figure 1: Extending the traditional Moreau-Broto atom–atom autocorrelation kernel to the atom–atom bond–bond bond–atom (AABBA) kernel of this work. TMC = transition metal complex; AC = autocorrelation; $\mathcal{G}$ = graph; p = atomic and bond properties; NBO = natural bond orbital.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

In this work, we introduce the atom–atom, bond–bond, bond–atom (AABBA) graph kernel, which extends the traditional Moreau-Broto atom–atom autocorrelations[44] with bond–bond and bond–atom terms. These terms include bond properties providing both geometric (*e.g.* bond distance) and electronic (*e.g.* bond order) structure information. Figure 1 illustrates this concept using a transition metal complex (TMC) as an example. The idea of extracting bond–bond and bond–atom relationships with the AABBA kernel was inspired by these two theoretical frameworks; 1) natural bond orbital (NBO) analysis, in which localized atomic (*i.e.* lone pairs and vacancies) and bond (*i.e.* 2- and 3-center non-bonding, bonding, and anti-bonding) orbitals interact with each other;[47] and 2) message-passing in graph neural networks, in which local and global chemical environments are learned by informing the atomic nodes with representations of their neighboring atoms and the bond edges connecting to them.[48,49] The vectors generated by the AABBA kernel were leveraged in ML models predicting TMC properties, including neural networks, gradient boosting machines, and Gaussian processes. In particular, we predicted the energy barriers and bond distances of the Vaska's complex dataset.[50] For both properties, the AABBA-based ML models achieved accuracies significantly higher than those obtained with a baseline kernel including only atom–atom terms. Further, quantitative measures of feature relevance showed that several bond–bond and bond–atom terms were among the most important in the ML predictions.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The AABBA graph kernel

We implemented the kernel in two different flavors: AABBA(I) and AABBA(II), as illustrated in Figure 2. Both act on molecular graphs, which, for organic molecules, can be easily generated from a string representation like SMILES or the *xyz* coordinates. For the benchmark of this study, which is based on TMCs, the robustness of these graph generation approaches is compromised by the complex nature of the metal–ligand bonds. This issue was addressed by using undirected natural quantum graphs (u-NatQG) derived from NBO analysis.[41] The u-NatQG representation yielded a robust orbital-based definition of the graph topology, in which the atomic nodes and the bond edges were attributed with either generic properties (*e.g.* atomic number and bond order) or specific NBO electronic properties (*e.g.* natural atomic charges and bond orbital symmetries).

In the AABBA(I) kernel, the atom–atom terms were computed using the traditional Moreau-Broto autocorrelation (AA-AC; Figure 2). The origin of the AA-AC can be either *full* (*i.e.* using all graph nodes as reference), or *metal-centered* (*i.e.* using only the metal node as reference). From the origin, all AA-AC terms can be extended from depth zero to an arbitrary maximum value. Figure 3 illustrates the depth concept in the metal-centered framework for the three terms of the AABBA kernels. Further, though the default arithmetic operation is the product, division, addition, and subtraction are also available. A similar approach was used for the computation of the bond–bond (BB-AC) and bond–atom (BA-AC) terms, for which we implemented these two methods: 1) for the metal-centered BB-AC, we defined a "super-bond" edge origin merging all metal–ligand bonds (Figure 8 in the Appendix) by either adding (BB-AC) or averaging ($\overline{\text{BB}}$-AC) their properties and, 2) for the BA-AC, we defined arithmetic operations consistent with the different dimensionality of the atomic and bond property vectors (Equation 18). The final representation yielded by the AABBA(I) kernel was built by concatenating the AA-AC, $\overline{\text{BB}}$-AC, and BA-AC terms (which can also be used independently) into a single vector. The dimensionality of this vector could be easily augmented by expanding these terms with different origins and operators.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

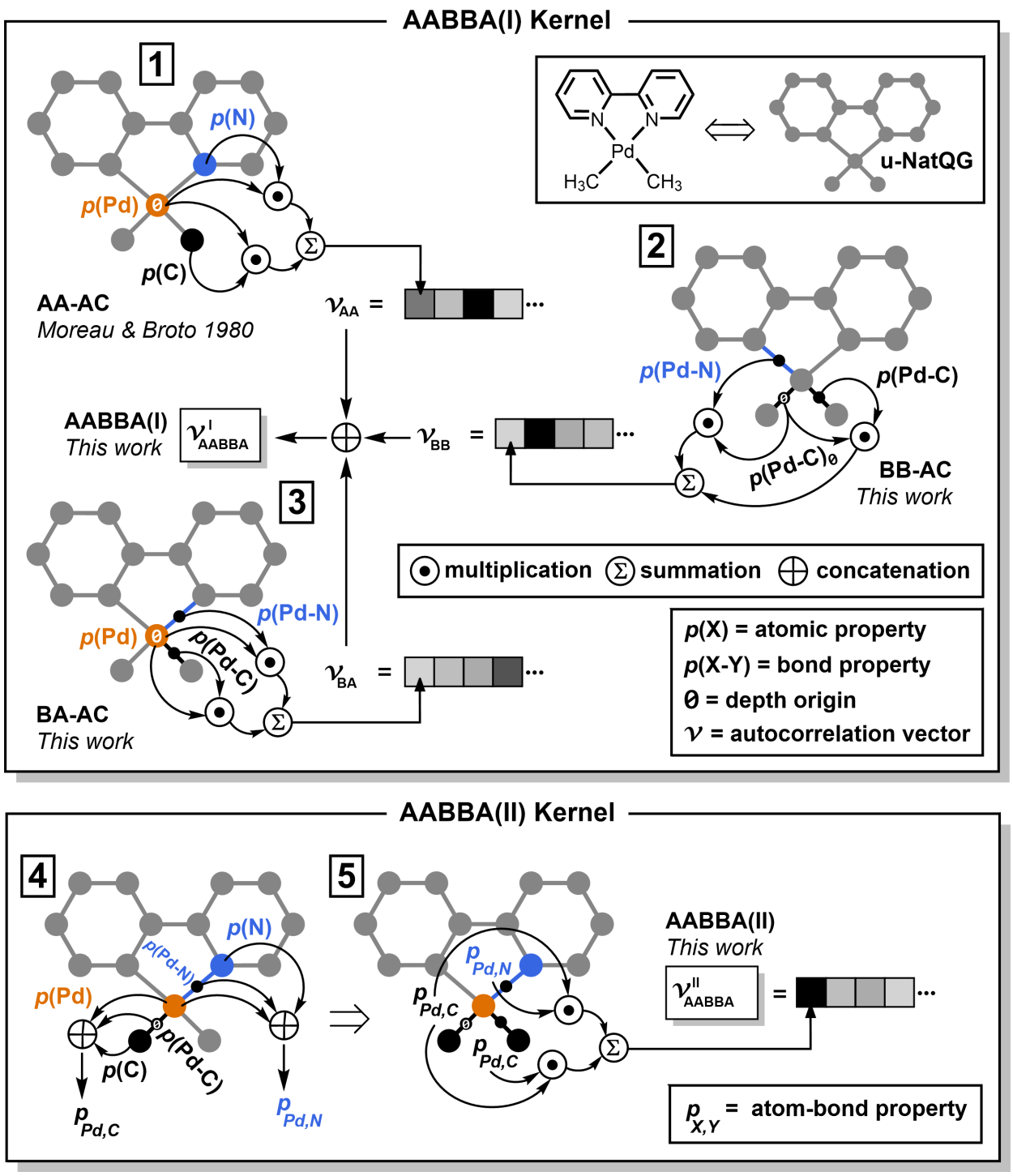

Figure 2: Computational graphs showing the autocorrelations of the AABBA kernels. For AABBA(I): The 1 Atom–Atom, 2 Bond–Bond, and 3 Bond–Atom autocorrelations. For AABBA(II): 4 Concatenation of atom and bond properties and 5 Autocorrelation of the resulting atom-bond properties. For the sake of clarity, only part of the arithmetic operations are shown at depths 0 and 1. **u-NatQG** = Undirected Natural Quantum Graph.[41]

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

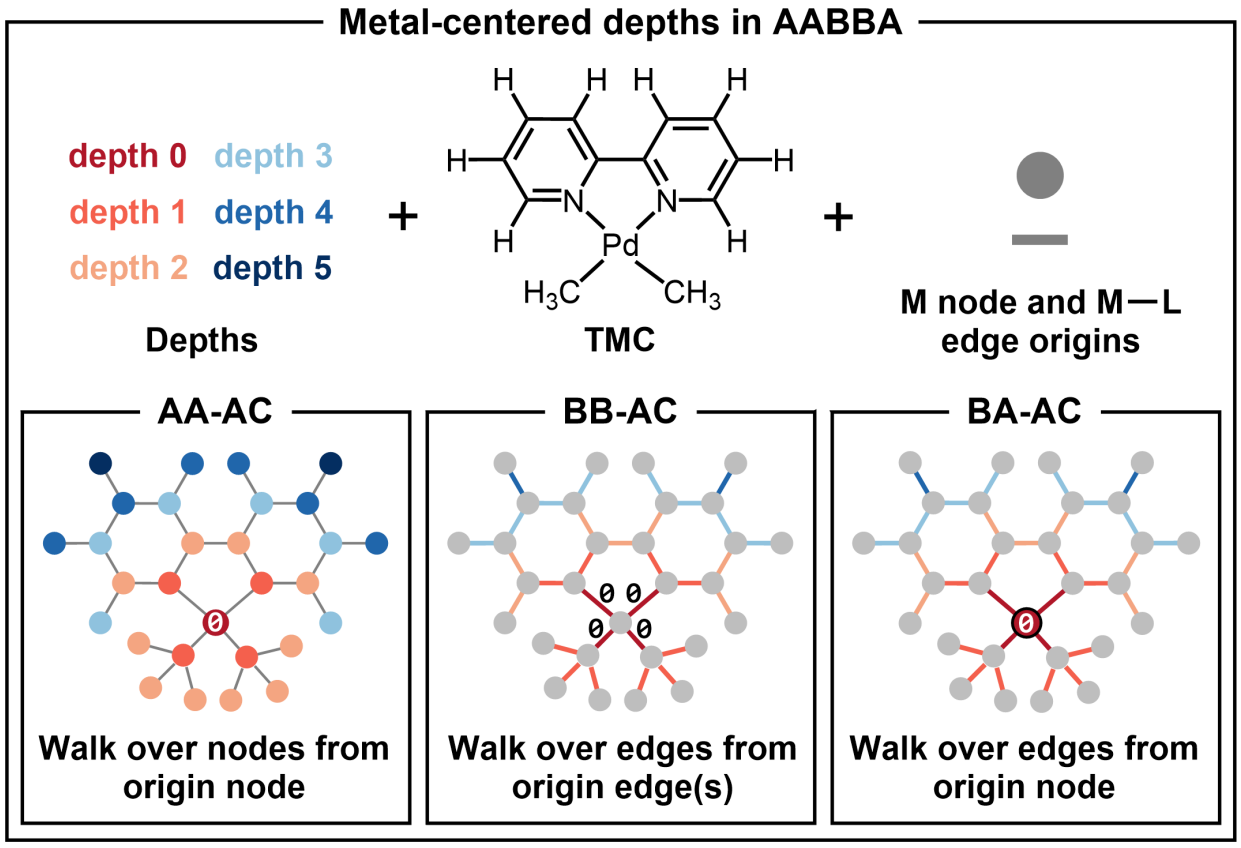

Figure 3: Metal-centered definition of the depth in the AABBA graph kernels. TMC = transition metal complex; M = metal; L = ligand; AC = autocorrelation; AA = atom–atom; BB = bond–bond; BA = bond–atom; **Ø** = depth origin.

In AABBA(II), selected atomic and bond properties associated with each atom—atom edge of the molecular graph were merged into an atom-bond property vector. For the generic properties, we defined three variants (AABBA(II)_{1-3}) differing in the definition of the electronegativity and geometry components (see Equation 28 in the Appendix). For the NBO properties, we defined two more variants (AABBA(II)_{4-5}), differing in the NBO data selected for the representation (Equations 29 and 30). Once defined, these merged property vectors were autocorrelated following the same procedure used for BB-AC in AABBA(I). In this case, there was no need to concatenate different terms since AABBA(II) generates the final fingerprint vector directly. This representation can be regarded as a dimensionality-reduced version of that generated by AABBA(I).

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

For a detailed description of the AABBA graph kernels, including the underlying equations, the dimensions of the resulting vectors, the definition of the metal-centered edge origin, and the systematic lists of generic and NBO properties, see the Appendix. Due to the modular nature of its vector concatenation operations, we believe that it should be easy to extend these kernels to materials made of distinct molecular blocks assembled in a structurally regular manner.

The Vaska's dataset

The Vaska's dataset used to benchmark the AABBA kernels is a curated collection of 1,947 iridium complexes with diverse σ-donor, σ/π-donor, and σ-donor/π-acceptor ligands. For each complex, the dataset provides computational results at the DFT(PBE/def2SVP) level for the transition state associated with the oxidative addition of molecular hydrogen, including the energy barrier and the breaking H$\cdots$H bond distance.

Systematic neural network models

The performance of the AABBA(I) and AABBA(II) graph kernels was first assessed with the Vaska's dataset using neural networks (NNs) for two regression tasks, one predicting the energy barrier and the other the H$\cdots$H bond distance. All models were based on a multilayer perceptron architecture with the following hyperparameters: Two hidden layers with 128 nodes each, ReLU activation, Adam optimizer minimizing the mean squared error (MSE) loss, and a training:validation:test data split of 80:10:10 (further details in the SI).

The autocorrelation vectors can be computed in multiple ways depending on the property set, arithmetic operator, origin, and maximum depth (see the Appendix). In the AABBA framework, this diversity is further expanded by the possibility of using these six graph kernels (Figure 2): AA-AC, BB-AC, $\overline{\text{BB}}$-AC, BA-AC, AABBA(I), and AABBA(II). As a first approach, we built the autocorrelation vectors in a systematic manner based on the results obtained by gradually increasing their complexity and dimensionality (Tables 1 and

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

2).

Table 1: Average and lowest test errors in the prediction of the Vaska's dataset energy barriers with neural networks. The inputs passed to the models were vectors defined with different graph kernels ($\mathcal{G}_\mathcal{K}$), property types (Prop), operators ($\hat{\text{Op}}$), origins ($\emptyset$), and maximum depths (D), yielding different dimensionality (*dim*). The mean absolute error (MAE) is given in kcal/mol.

Model	Input						Average Error[a]		Lowest Error[a]	
	$\mathcal{G}_\mathcal{K}$	Prop	$\hat{\text{Op}}$	$\emptyset$[b]	D	*dim*[c]	MAE	r^2	MAE	r^2
1	**AA**	**P**[d]	$\odot$[e]	**MC**	**3**	**18**	**1.22 ± 0.02**	**0.844 ± 0.007**	**1.16**	**0.850**
2	BB	P	$\odot$	MC	3	12	1.42 ± 0.03	0.801 ± 0.005	1.37	0.803
3	$\overline{\textbf{BB}}$	**P**	$\odot$	**MC**	**3**	**10**	**1.40 ± 0.04**	**0.765 ± 0.016**	**1.26**	**0.791**
4	BA	P	$\odot$	MC	3	20	2.14 ± 0.04	0.571 ± 0.013	2.05	0.595
5	**I**[f]	**P**	$\odot$	**MC**	**3**	**48**	**0.90 ± 0.02**	**0.914 ± 0.003**	**0.86**	**0.916**
6	I	P	$\ominus_\chi$[e]	MC	3	47	0.91 ± 0.01	0.913 ± 0.003	0.89	0.919
7	I	P	$\oslash_R$[e]	MC	3	47	0.92 ± 0.02	0.911 ± 0.004	0.89	0.919
8	I	NBO[g]	$\odot$	MC	3	212	0.89 ± 0.02	0.908 ± 0.007	0.85	0.913
9	**I**	**NBO**	$\odot$	**F**	**3**	**220**	**0.79 ± 0.02**	**0.927 ± 0.002**	**0.76**	**0.928**
10	I	NBO	$\odot$	MC	4	263	0.90 ± 0.02	0.899 ± 0.008	0.84	0.925
11	I	NBO	$\odot$	MC	5	298	0.90 ± 0.02	0.904 ± 0.007	0.87	0.899
12	**I**	**NBO**	$\odot$	**MC**	**6**	**303**	**0.88 ± 0.02**	**0.907 ± 0.007**	**0.84**	**0.918**
13	I[h]	NBO	$\odot$	F	3	223	0.78 ± 0.02	0.928 ± 0.004	0.73	0.933
14	$\textbf{II}_1$[f]	**P**	$\odot$	**MC**	**3**	**33**	**0.94 ± 0.02**	**0.906 ± 0.002**	**0.89**	**0.913**
15	II_2	P	$\odot$	MC	3	29	0.96 ± 0.03	0.904 ± 0.005	0.86	0.917
16	II_3	P	$\odot$	MC	3	33	0.94 ± 0.02	0.908 ± 0.004	0.90	0.913
17	II_4	NBO	$\odot$	MC	3	92	0.86 ± 0.02	0.918 ± 0.003	0.82	0.926
18	II_5	NBO	$\odot$	MC	3	80	0.94 ± 0.02	0.907 ± 0.004	0.88	0.915
19	II_4	NBO	$\odot$	F	3	98	1.15 ± 0.03	0.850 ± 0.007	1.05	0.862
20	II_4	NBO	$\odot$	MC	4	111	0.88 ± 0.02	0.916 ± 0.005	0.85	0.914
21	$\textbf{II}_4$	**NBO**	$\odot$	**MC**	**5**	**129**	**0.85 ± 0.03**	**0.916 ± 0.004**	**0.81**	**0.923**
22	II_4	NBO	$\odot$	MC	6	129	0.85 ± 0.01	0.919 ± 0.003	0.83	0.921

[a]From ten repetitions with a training:validation:test split of 80:10:10; [b]Metal-centered (MC) or full (F); [c]After removing redundant dimensions; [d]*I.e.* P_A, P_B, and P_AB periodic and generic property sets; [e]All properties correlated by product ($\odot$), except the electronegativity in entry 6 (subtracted, $\ominus_\chi$), and the covalent radius in entry 7 (divided, $\oslash_R$); [f]Entries 5-13 and 14-22 correspond to the AABBA(I) and AABBA(II) kernels; [g]*I.e.* $\text{P}_\text{A,NBO}$, $\text{P}_\text{B,NBO}$, and $\text{P}_\text{AB,NBO}$ NBO property sets. [h]Including whole-graph properties.

See the Appendix for further details.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The results collected in Table 1 for the prediction of the oxidative addition barrier show that the Moreau-Broto AA-AC achieved a mean absolute error (MAE) of 1.16 kcal/mol, using the metal-centered autocorrelation of generic properties at a maximum depth of three (model 1). Following the same approach, we also tested the autocorrelations leveraging bond properties, *i.e.* the BB-, $\overline{\text{BB}}$-, and BA-AC. In line with the smaller amount of properties describing the bonds (three) relative to the atoms (nine), these autocorrelations showed a poorer performance, though $\overline{\text{BB}}$-AC achieved a remarkable accuracy, with MAE = 1.26 kcal/mol (model 3). Further, when the associated autocorrelation vector was concatenated with those extracted by the AA- and BA-AC kernels, the resulting AABBA(I) representation gave a MAE of 0.86 kcal/mol (model 5), which was 26% smaller than that of the AA-AC baseline.

Going further with the AABBA(I) kernel, we investigated how accuracy could be improved by fine-tuning other parameters. First, we considered other property operators and, in particular, the use of deltametric (property differences) and ratiometric (property ratios) ACs for the electronegativity and covalent radius, respectively (models 6 and 7 in Table 1). These operators encoded local variations in bond polarization and relative atomic size but, in practice, neither of them yielded lower MAEs. Next, we replaced the generic properties with the NBO, which, with the full origin, yielded the second lowest MAE in the series of numerical experiments: 0.76 kcal/mol (model 9). With the metal-centered origin, which is more useful from the perspective of explainability, the MAE could be reduced to 0.84 kcal/mol after increasing the maximum depth of the representation to six (model 12). In a final experiment, we enriched the most accurate representation by adding whole-graph properties but the increase in accuracy was very small (model 13).

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

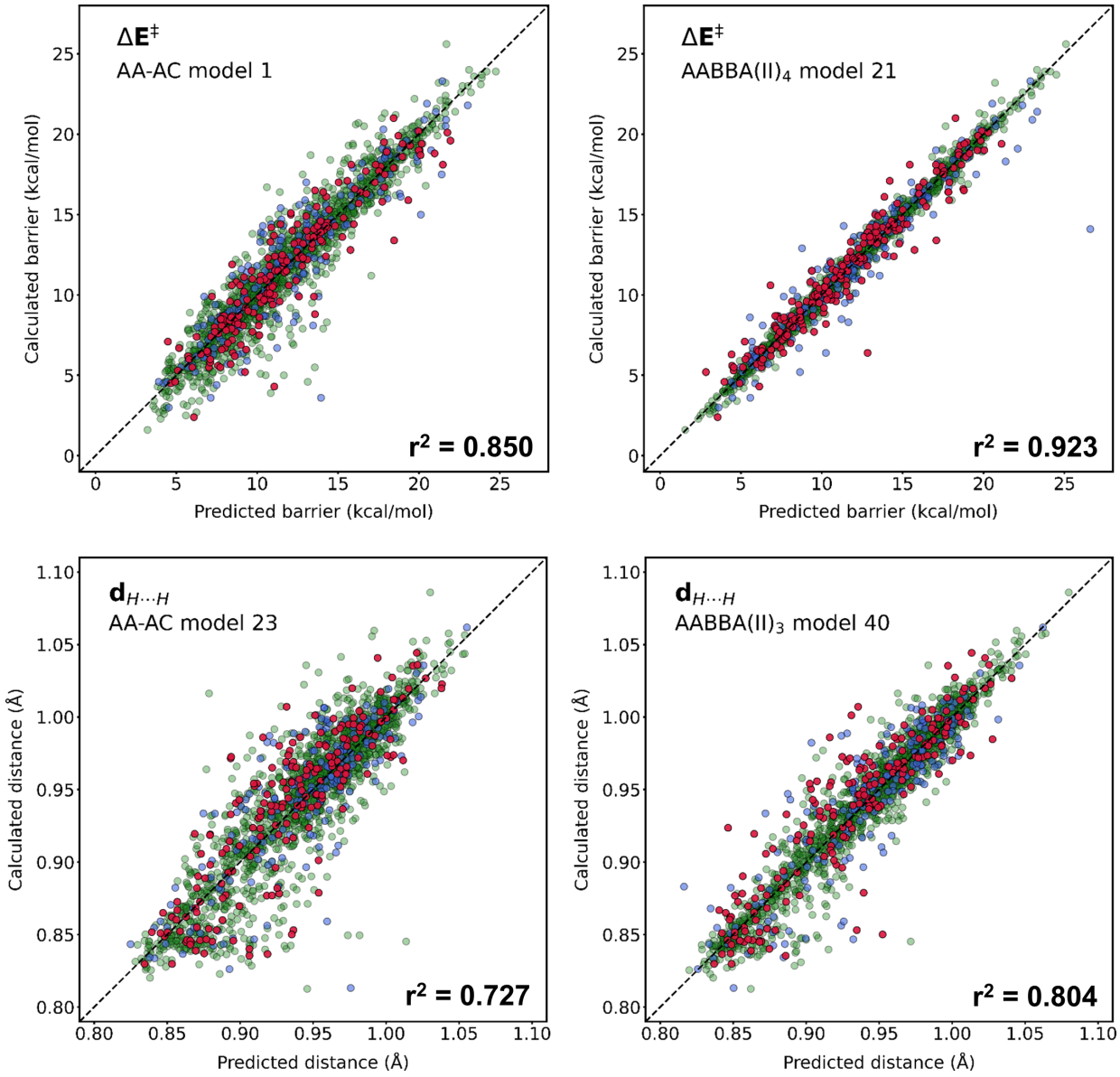

Figure 4: Pair plots showing the correlation between the DFT-calculated and NN-predicted energy barrier ($\Delta E^{\ddagger}$) and breaking H$\cdots$H bond distance ($d_{H\cdots H}$) of the Vaska's dataset. All models refer to Tables 1 and 2. Data points color code: • training, • validation, and • test.

The AABBA(II) kernel also outperformed the AA-AC baseline, reducing the MAE to an extent similar to AABBA(I), from 1.16 to 0.89 kcal/mol, using the AABBA(II)$_1$ kernel (model 14 in Table 1). In contrast with AABBA(I), the use of NBO properties gave the highest accuracy when the AABBA(II)$_4$ kernel was combined with the metal-centered origin and a maximum depth of five, yielding a MAE of 0.81 kcal/mol (model 21). Though it did not achieve the lowest MAE of the series, a significant advantage of this representation is

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

its low dimensionality (129) relative to its AABBA(I) equivalent (298; model 11). The pair plots in Figure 4 show the performance of this AABBA(II) model relative to the AA-AC baseline.

The prediction of the H$\cdots$H bond distance was explored following the same systematic approach (Table 2). Using also the metal-centered autocorrelation of the generic properties at a maximum depth of three, the graph kernels including only atom–atom or bond–bond terms, *i.e.* AA-, $\overline{\text{BB}}$-, and BB-AC, yielded similar accuracies, with MAE differences smaller than 0.002 Å (models 23-25). In contrast with the prediction of the energy barriers (Table 1), where AA-AC was the most accurate of these three kernels, the lowest MAE was hereby achieved with $\overline{\text{BB}}$-AC ($2.09{\cdot}10^{-2}$ Å; model 25). This observation aligns with the notion that bond properties are crucial in predicting the distance, thereby contributing to improved results. The BA-AC kernel was also tested but it exhibited poorer performance (model 26).

Keeping the same properties, operator, origin, and depth, and expanding the autocorrelations with the AABBA(I) kernel, the MAE was reduced to $1.93{\cdot}10^{-2}$ Å (model 27 in Table 2). Interestingly, by replacing the product autocorrelation of the electronegativity with its deltametric, the MAE was further minimized down to $1.87{\cdot}10^{-2}$ Å (model 30), which was the lowest achieved with this kernel, showing the value of encoding bond polarity with electronegativity differences. Next, we moved to NBO properties, which, unlike the prediction of the energy barriers (Table 1), did not improve the results. The model closest in accuracy (35) used full autocorrelations with a maximum depth of five, thus producing large input vectors ($dim = 316$).

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Table 2: Average and lowest test errors in the prediction of the Vaska's dataset H···H distance with neural networks. The inputs passed to the models were vectors defined with different graph kernels ($\mathcal{G}_\mathcal{K}$), property types (Prop), operators ($\hat{\text{Op}}$), origins ($\emptyset$), and maximum depths (D), yielding different dimensionality (*dim*). The mean absolute error (MAE) is given in Å.

Model	Input						Average Error[a]		Lowest Error[a]	
	$\mathcal{G}_\mathcal{K}$	Prop	$\hat{\text{Op}}$	$\emptyset$[b]	D	*dim*[c]	MAE	r^2	MAE	r^2
23	**AA**	**P**[d]	$\odot$[e]	**MC**	**3**	**18**	$\mathbf{2.38{\cdot}10^{-2} \pm 9{\cdot}10^{-4}}$	**0.687 ± 0.012**	$\mathbf{2.11{\cdot}10^{-2}}$	**0.727**
24	BB	P	$\odot$	MC	3	10	$2.30{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.706 ± 0.007	$2.21{\cdot}10^{-2}$	0.714
25	$\overline{\textbf{BB}}$	**P**	$\odot$	**MC**	**3**	**12**	$\mathbf{2.35{\cdot}10^{-2} \pm 1.0{\cdot}10^{-3}}$	**0.673 ± 0.020**	$\mathbf{2.09{\cdot}10^{-2}}$	**0.729**
26	BA	P	$\odot$	MC	3	20	$3.03{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.537 ± 0.012	$2.91{\cdot}10^{-2}$	0.551
27	**I**[f]	**P**	$\odot$	**MC**	**3**	**48**	$\mathbf{2.07{\cdot}10^{-2} \pm 4{\cdot}10^{-4}}$	**0.747 ± 0.006**	$\mathbf{1.93{\cdot}10^{-2}}$	**0.767**
28	I	P	$\odot$	F	3	220	$2.12{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$	0.716 ± 0.015	$1.98{\cdot}10^{-2}$	0.739
29	I	P	$\odot$	F[g]	3	220	$2.54{\cdot}10^{-2} \pm 1.0{\cdot}10^{-3}$	0.632 ± 0.019	$2.37{\cdot}10^{-2}$	0.669
30	**I**	**P**	$\ominus_\chi$[e]	**MC**	**3**	**47**	$\mathbf{1.96{\cdot}10^{-2} \pm 5{\cdot}10^{-4}}$	**0.767 ± 0.007**	$\mathbf{1.87{\cdot}10^{-2}}$	**0.769**
31	I	P	$\oslash_R$[e]	MC	3	47	$2.04{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.764 ± 0.004	$1.96{\cdot}10^{-2}$	0.761
32	I	NBO[h]	$\odot$	MC	3	212	$2.27{\cdot}10^{-2} \pm 8{\cdot}10^{-4}$	0.663 ± 0.021	$2.13{\cdot}10^{-2}$	0.702
33	I	NBO	$\odot$	MC	4	263	$2.22{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$	0.676 ± 0.012	$2.12{\cdot}10^{-2}$	0.690
34	I	NBO	$\odot$	MC	5	298	$2.16{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.705 ± 0.020	$2.04{\cdot}10^{-2}$	0.706
35	**I**	**NBO**	$\odot$	**F**	**5**	**316**	$\mathbf{2.10{\cdot}10^{-2} \pm 1.0{\cdot}10^{-3}}$	**0.718 ± 0.021**	$\mathbf{1.89{\cdot}10^{-2}}$	**0.748**
36	I	NBO	$\odot$	F[g]	5	316	$2.60{\cdot}10^{-2} \pm 5.2{\cdot}10^{-3}$	0.586 ± 0.153	$2.06{\cdot}10^{-2}$	0.727
37	I	NBO	$\odot$	MC	6	303	$2.17{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$	0.704 ± 0.016	$1.99{\cdot}10^{-2}$	0.745
38	II_1[f]	P	$\odot$	MC	3	33	$1.91{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.771 ± 0.009	$1.81{\cdot}10^{-2}$	0.788
39	II_2	P	$\odot$	MC	3	29	$1.93{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.768 ± 0.006	$1.81{\cdot}10^{-2}$	0.772
40	$\mathbf{II}_3$	**P**	$\odot$	**MC**	**3**	**33**	$\mathbf{1.85{\cdot}10^{-2} \pm 5{\cdot}10^{-4}}$	**0.781 ± 0.009**	$\mathbf{1.72{\cdot}10^{-2}}$	**0.804**
41	II_3	P	$\odot$	F	3	54	$1.97{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$	0.747 ± 0.013	$1.87{\cdot}10^{-2}$	0.765
42	II_4	NBO	$\odot$	MC	3	92	$2.21{\cdot}10^{-2} \pm 5{\cdot}10^{-4}$	0.685 ± 0.013	$2.13{\cdot}10^{-2}$	0.704
43	II_5	NBO	$\odot$	MC	3	80	$2.20{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.687 ± 0.024	$2.05{\cdot}10^{-2}$	0.749
44	II_3	P	$\odot$	MC	4	42	$1.87{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.776 ± 0.009	$1.82{\cdot}10^{-2}$	0.785
45	II_3	P	$\odot$	MC	5	51	$1.86{\cdot}10^{-2} \pm 5{\cdot}10^{-4}$	0.766 ± 0.011	$1.78{\cdot}10^{-2}$	0.795
46	II_3	P	$\odot$	MC	6	51	$1.86{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.770 ± 0.008	$1.74{\cdot}10^{-2}$	0.792
47	II_3[i]	P	$\odot$	MC	3	36	$1.86{\cdot}10^{-2} \pm 4{\cdot}10^{-4}$	0.777 ± 0.008	$1.74{\cdot}10^{-2}$	0.799

[a]From ten repetitions with a training:validation:test split of 80:10:10; [b]Metal-centered (MC) or full (F); [c]After removing redundant dimensions; [d]*I.e.* P_A, P_B, and P_AB periodic and generic property sets; [e]All properties correlated by product ($\odot$), except the electronegativity in entry 8 (subtracted, $\ominus_\chi$), and the covalent radius in entry 9 (divided, $\oslash_R$); [f]Entries 5-15 and 16-24 correspond to the AABBA(I) and AABBA(II) kernels, respectively; [g]From an extended neural network of 3 hidden layers with 256 nodes each; [h]*I.e.* $\text{P}_\text{A,NBO}$, $\text{P}_\text{B,NBO}$, and $\text{P}_\text{AB,NBO}$ NBO property sets. [i]Also including whole-graph properties. See the Appendix for further details.

The lower-dimensionality representations extracted with the AABBA(II) kernel also showed better performance with the generic properties than with the NBO (Table 2). The most accurate model in this series yielded MAE = $1.72 \cdot 10^{-2}$ Å (model 40) using the AABBA(II)_3 representation, which combines electronegativity differences with covalent radius instead of optimized interatomic distances, being thus geometry-agnostic. AABBA(II)_3 reduced the MAE of the AA-AC baseline by 18% (the performance of both models is compared in Figure 4). Increasing the maximum depth to six (models 44-46) or extending the representation with whole-graph properties (model 47) did not improve accuracy any further, suggesting that the H···H distance is dominated by local rather than global effects.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Dimensionality reduction by feature relevance

The results in Tables 1 and 2 show that the performance of the NN models is sensitive to the nature of the graph kernel, as well as parameters like the origin and depth of the representations. The manual adjustment of these variables is challenging and, beyond domain knowledge and heuristics like, for example, using larger depths to capture remoter effects, optimal solutions could be easily missed. From this perspective, the comparison of the results obtained with the AABBA(I) and AABBA(II) kernels suggested that dimensionality reduction could be an appropriate strategy for tackling this issue.

Gradient boosting machines

We explored feature selection based on ensemble models by defining autocorrelation vectors with maximal dimensionality (MD), which were used to train gradient boosting machine (GBM) models predicting the Vaska's dataset energy barriers and breaking H$\cdots$H bond distances. The GBM results were analyzed with the double aim of 1) comparing the different autocorrelation features from the perspective of their importance in the predictions, and 2) making a selection of these features for dimensionality-reduced ML models, which are presented in the next two sections.

The length of the MD autocorrelation vectors was maximized with the AABBA(I) graph kernel (Figure 2), using both the full and metal-centered origins, with a maximum depth of six, and all four arithmetic operators (see the Appendix). All these terms were concatenated to form the final $v^{I,MD}_{AABBA}$ representation, which was either an 671- or 2750-dimensional vector, depending on whether the generic or NBO properties, respectively, were used to compute the autocorrelations. These dimensionalities are between six- and nine-times larger than those of the vectors fed to the NN models in Tables 1 and 2.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Table 3: Average and lowest test errors in the prediction of the Vaska's dataset energy barrier ($\Delta E^{\ddagger}$) and H$\cdots$H distance ($d_{H\cdots H}$) with GBM models. The inputs passed to the models were $v^{I,MD}_{AABBA}$ vectors defined with different property types (Prop) yielding different (maximal) dimensionalities (*dim*). The mean absolute errors (MAEs) are given in kcal/mol for the barriers and Å for the distances.

Target	Input			Average Error[a]		Lowest Error[a]	
	$\mathcal{G}_K$	Prop	*dim*[b]	MAE	r^2	MAE	r^2
$\Delta E^{\ddagger}$	**I**	**P[c]**	**671**	**1.00 ± 0.05**	**0.891 ± 0.004**	**0.91**	**0.919**
$\Delta E^{\ddagger}$	AA	P[c]	223	1.09 ± 0.05	0.869 ± 0.02	1.01	0.896
$\Delta E^{\ddagger}$	BB	P[c]	188	1.20 ± 0.04	0.858 ± 0.016	1.14	0.860
$\Delta E^{\ddagger}$	BA	P[c]	260	1.29 ± 0.02	0.840 ± 0.018	1.24	0.867
$d_{H\cdots H}$	**I**	**P[c]**	**671**	**$1.86{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$**	**0.738 ± 0.022**	**$1.73{\cdot}10^{-2}$**	**0.760**
$d_{H\cdots H}$	AA	P[c]	223	$2.00{\cdot}10^{-2} \pm 5{\cdot}10^{-4}$	0.711 ± 0.021	$1.89{\cdot}10^{-2}$	0.750
$d_{H\cdots H}$	BB	P[c]	188	$2.06{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.706 ± 0.021	$1.97{\cdot}10^{-2}$	0.732
$d_{H\cdots H}$	AB	P[c]	260	$2.16{\cdot}10^{-2} \pm 7{\cdot}10^{-4}$	0.671 ± 0.025	$2.05{\cdot}10^{-2}$	0.703
$\Delta E^{\ddagger}$	**I**	**NBO[d]**	**2750**	**0.84 ± 0.04**	**0.920 ± 0.016**	**0.77**	**0.940**
$\Delta E^{\ddagger}$	AA	NBO[d]	680	0.85 ± 0.04	0.918 ± 0.019	0.79	0.942
$\Delta E^{\ddagger}$	BB	NBO[d]	1068	1.05 ± 0.02	0.888 ± 0.012	1.02	0.902
$\Delta E^{\ddagger}$	BA	NBO[d]	1002	1.07 ± 0.04	0.877 ± 0.018	0.98	0.902
$d_{H\cdots H}$	**I**	**NBO[d]**	**2750**	**$1.79{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$**	**0.753 ± 0.018**	**$1.67{\cdot}10^{-2}$**	**0.784**
$d_{H\cdots H}$	AA	NBO[d]	680	$1.80{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.748 ± 0.022	$1.71{\cdot}10^{-2}$	0.777
$d_{H\cdots H}$	BB	NBO[d]	1068	$1.99{\cdot}10^{-2} \pm 9{\cdot}10^{-4}$	0.711 ± 0.032	$1.86{\cdot}10^{-2}$	0.762
$d_{H\cdots H}$	BA	NBO[d]	1002	$1.95{\cdot}10^{-2} \pm 6{\cdot}10^{-4}$	0.708 ± 0.025	$1.85{\cdot}10^{-2}$	0.739

[a]From 5-fold cross-validation; [b]After removing redundant dimensions; [c]*I.e.* P_A, P_B, and P_{AB} generic property sets; [d]*I.e.* $P_{A,NBO}$, $P_{B,NBO}$, and $P_{AB,NBO}$ NBO property sets. See the Appendix for further details.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Regression trees (RTs) were used as base learners in the GBM ensemble models, which included 1000 RTs with a maximum depth of 5. The models were trained by minimizing the MSE loss with a learning rate of 0.05, and were tested by 5-fold cross-validation (further details in the SI). Table 3 shows the performance of the GBMs in the prediction of the Vaska's energy barriers and H$\cdots$H distances. For both regression tasks, the NBO-informed representations gave higher accuracies than those based on generic properties. Another interesting observation is that the full $v^{I,MD}_{AABBA}$ vector achieved higher accuracies than the $\overline{\text{BB}}$, BA, and AA alone, though the latter performed at nearly the same level when using NBO properties. The most accurate models yielded remarkably low errors; *i.e.* 0.77 kcal/mol for the barrier ($r^2 = 0.940$) and $1.67{\cdot}10^{-2}$ Å for the distance ($r^2 = 0.784$). The latter model was the most accurate of the series reported in this work for the prediction of distances with NBO properties. Since GBMs are designed to maximize accuracy by selecting the most relevant features, these results suggested that the many dimensions of these autocorrelation vectors could be redundant to a significant extent, thus showing that dimensionality reduction would likely be an efficient strategy.

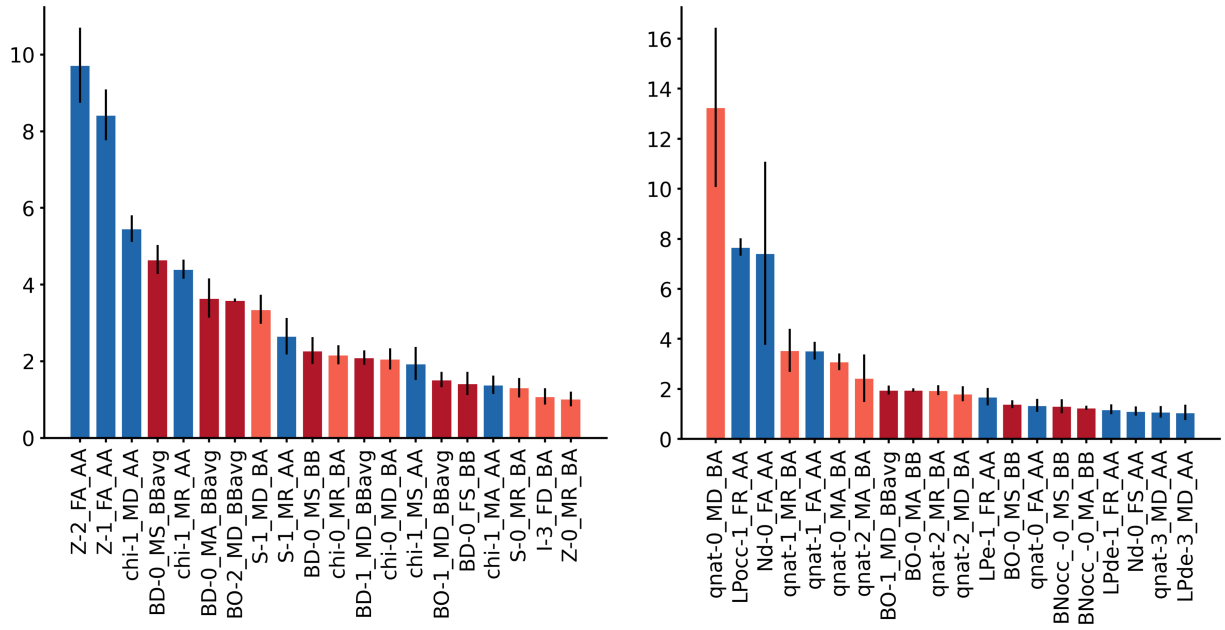

Figure 5: Relevances (y-axes, in %) of the twenty most important AABBA components (x-axes) from the GBM models predicting the Vaska's energy barriers. Bar color code: ■ AA-AC, ■ BB-AC and $\overline{\text{BB}}$-AC, and ■ BA-AC.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

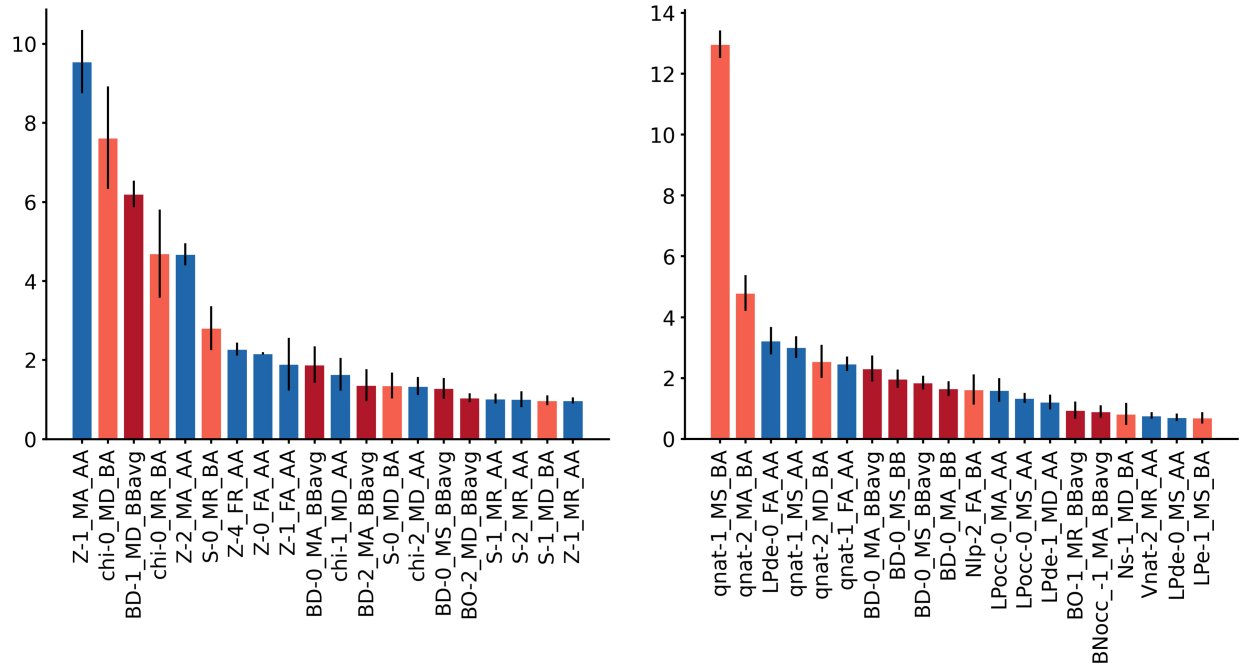

Figure 6: Relevances (y-axes, in %) of the twenty most important AABBA components (x-axes) from the GBM models predicting the Vaska's breaking H$\cdots$H distances. Bar color code: ■ AA-AC, ■ BB-AC and $\overline{\text{BB}}$-AC, and ■ BA-AC.

The relevance of the features, *i.e.* the components of the $v^{I,MD}_{AABBA}$ autocorrelation vectors, was computed using the Friedman MSE criterion. This criterion exploits the ensemble nature of the GBMs, quantifying the reduction of the residual sum of squares of the model by any given feature relative to the total. Figures 5 and 6 show the twenty most relevant features in the GBM models predicting the barriers and distances, respectively, from the generic and NBO representations. In all cases, the three atom–atom, bond–bond, and bond–atom components of the AABBA(I) graph kernel yielded features of high relevance. For example, in the prediction of the energy barrier with generic properties, there is a similar number of AA, BB, and BA features (*i.e.* 7, 7, and 6, respectively) among the most relevant, which is also true in the other three cases. Table 4, which collects and describes only the 5 most relevant features, also shows some interesting trends. For the barriers, there is a mix of full and metal-centered features, which, with the generic properties, refer to chemical composition (Z) and electronegativity (χ), whereas, with the NBO properties, refer to natural electron counts and charges (q_{Nat}). There is also a mix of full and metal-centered features, which

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

is consistent with the influence of both local and global effects on the oxidative addition barrier. Further, features like the bond–atom metal-centered charges can be directly related to the critical role played by the electron density of the metal center. For the distances, most features are metal-centered at depth zero or one, in line with the local nature of the H$\cdots$H bond cleavage, which takes place within the first coordination sphere of the metal center. As for the barrier, the generic features involve mostly Z and χ, whereas the NBO are dominated by q_{Nat}. Regardless of the target, the product autocorrelation operator is the most common, appearing in half of the features, whereas the other half is diverse, including the ratiometric, summetric, and deltametric operators.

Table 4: Five most relevant features in the prediction of the Vaska's dataset energy barrier ($\Delta E^{\ddagger}$) and breaking H$\cdots$H distance ($d_{H\cdots H}$) with generic (P) and natural bond orbital (NBO) properties. The rank refers to feature relevance as derived from the GBM models.

Target = $\Delta E^{\ddagger}$, Properties = P		
Rank	Feature label	Feature description
1	Z-2_FA_AA	Full atom–atom autocorrelation of atomic number at depth 2
2	Z-1_FA_AA	Full atom–atom autocorrelation of atomic number at depth 1
3	χ-1_MD_AA	Metal-centered atom–atom deltametric of electronegativity at depth 1
4	BD-0_MS_$\overline{BB}$	Metal-centered averaged bond–bond summetric of distance at depth 0
5	χ-1_MR_AA	Metal-centered atom–atom ratiometric of electronegativity at depth 1
Target = $\Delta E^{\ddagger}$, Properties = NBO		
Rank	Feature label	Feature description
1	q_{Nat}-0_MD_BA	Metal-centered bond–atom deltametric of natural charge at depth 0
2	LP_{Occ}-1_FR_AA	Full atom–atom ratiometric of lone-pair occupancies at depth 1
3	N_d-0_FA_AA	Full atom–atom autocorrelation of d-electron count at depth 0
4	q_{Nat}-1_MR_BA	Metal-centered bond–atom ratiometric of natural charge at depth 1
5	q_{Nat}-1_FA_AA	Full atom–atom autocorrelation of natural charge at depth 1
Target = $d_{H\cdots H}$, Properties = P		
Rank	Feature label	Feature description
1	Z-1_MA_AA	Metal-centered atom–atom autocorrelation of atomic number at depth 1
2	χ-0_MD_BA	Metal-centered bond–atom deltametric of electronegativity at depth 0
3	BD-1_MD_$\overline{BB}$	Metal-centered averaged bond–bond deltametric of distance at depth 0
4	χ-0_MR_BA	Metal-centered bond–atom ratiometric of electronegativity at depth 0
5	Z-2_MA_AA	Metal-centered atom–atom autocorrelation of atomic number at depth 2
Target = $d_{H\cdots H}$, Properties = NBO		
Rank	Feature label	Feature description
1	q_{Nat}-1_MS_BA	Metal-centered bond–atom summetric of natural charge at depth 1
2	q_{Nat}-2_MA_BA	Metal-centered bond–atom autocorrelation of natural charge at depth 2
3	$LP_{\Delta E}$-0_FA_AA	Full atom–atom autocorrelation of lone-pair energy gap at depth 0
4	q_{Nat}-1_MS_AA	Metal-centered atom–atom summetric of natural charge at depth 1
5	q_{Nat}-2_MD_BA	Metal-centered bond–atom deltametric of natural charge at depth 2

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Gaussian processes

The use of AABBA representations of reduced dimensionality was investigated with Gaussian processes (GP) predicting the Vaska's energy barriers and breaking H$\cdots$H distances. These models were based on a composed Linear-RBF kernel ($\mathcal{K}_{LR}$; RBF = radial basis function) in which both covariance functions were multiplied; *i.e.*

$$\mathcal{K}_{LR}(x, x') = \sigma^2 \cdot x^T x' \cdot \exp(-\frac{1}{2}(x - x')^T \lambda^{-2}(x - x')) \quad (1)$$

where (x, x') is a pair of data points, σ^2 is the variance, and λ is the length-scale of the kernel. Preliminary studies showed that these two kernels perform at a lower level when used separately in these regression tasks.

The $v^{I,MD}_{AABBA}$ vectors with maximal dimensionality were simplified by gradually removing terms according to the accumulated relevance found with the GBM models (*vide supra*). MAEs were computed for each pruned representation, taking that of the full $v^{I,MD}_{AABBA}$ vector as the baseline. In the prediction of $\Delta E^{\ddagger}$ with generic properties (Figure 7), we observed a gentle decrease of the MAE until a reduced dimensionality of ~50, followed by a steep increase of the MAE at smaller dimensionalities. The top-performance model used a 46-dimensional AABBA representation (80% accumulated relevance), in which 59% of the terms were either BB or BA (Table 5) – Remarkably, with MAE = 0.67 kcal/mol and r^2 = 0.947, this GP model was the most accurate of the series reported in this work. With the NBO features, the convergence of the MAE with the reduction of the dimensionality was less stable and two minima were observed, as clearly shown by the accumulated relevance plots (Figure S7). The lowest MAE model was based on a 115-dimensional AABBA vector (86% accumulated relevance) in which the BB and BA terms amounted 54% of the total dimensions. With MAE = 0.77 kcal/mol and $r^2 = 0.930$, this model performed at a level similar to that of the most accurate NNs found in the systematic study (*e.g.* model 9 in Table 1).

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

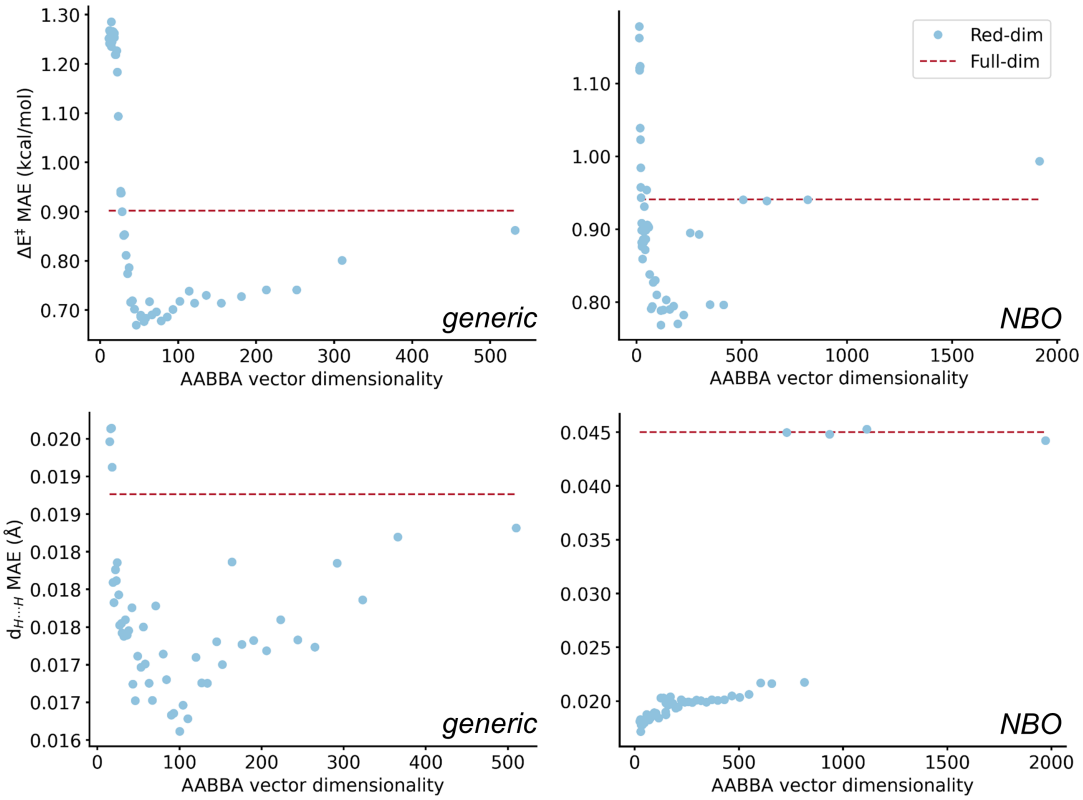

Figure 7: Influence of reducing the dimensionality of the AABBA representation in the MAEs of the GP predicting the Vaska's barriers ($\Delta E^{\ddagger}$; top) and distances ($d_{H\cdots H}$; bottom) from generic (left) and NBO (right) properties. The legend applies to all four plots.

Table 5: Test errors in the prediction of the Vaska's dataset energy barrier ($\Delta E^{\ddagger}$) and H$\cdots$H distance ($d_{H\cdots H}$) with Gaussian processes. The inputs passed to the models were $v^{I,MD}_{AABBA}$ vectors defined with different property types (Prop) after being pruned to a reduced dimensionality (*rdim*) based on the accumulated relevance (AR) found with the GBM models. The mean absolute errors (MAEs) are given in kcal/mol for the barriers and Å for the distances.

Target	Input				Error[a]	
	Prop	*rdim*[b]	AR	AA:BB:BA	MAE	r^2
$\Delta E^{\ddagger}$	P[c]	46	80%	19:12:15	0.67	0.947
$\Delta E^{\ddagger}$	P[c]	46	80%	19:12:15	1.32	0.799
$\Delta E^{\ddagger}$	NBO[c]	115	86%	53:29:33	0.77	0.930
$\Delta E^{\ddagger}$	NBO[c]	115	86%	53:29:33	1.14	0.845
$d_{H\cdots H}$	P[c]	100	82%	42:22:36	$1.61\cdot10^{-2}$	0.824
$d_{H\cdots H}$	P[c]	100	82%	42:22:36	$2.42\cdot10^{-2}$	0.571
$d_{H\cdots H}$	NBO[c]	28	52%	13:7:8	$1.72\cdot10^{-2}$	0.783
$d_{H\cdots H}$	NBO[c]	28	52%	13:7:8	$2.42\cdot10^{-2}$	0.582

[a]From ten repetitions with a training:validation:test split of 80:10:10 (black) or 20:40:40 (purple); [b]After pruning the AABBA representation with respect to the GBM relevances; [c]*I.e.* P_A, P_B, and P_{AB} generic property sets; [d]*I.e.* $P_{A,NBO}$, $P_{B,NBO}$, and $P_{AB,NBO}$ NBO property sets. See the Appendix for further details.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

In the prediction of the distances (Figure 7), the generic properties showed a trend that was less stable and yet similar to that observed for the barriers. The MAE was minimized to $1.61 \cdot 10^{-2}$ Å ($r^2 = 0.824$) with a simplified input of 100 dimensions (82% accumulated relevance), of which 58% were either BB or BA (Table 5). This model was also the most accurate of the series, thus showing that the use of Gaussian processes with reduced AABBA representations computed from generic properties is a powerful approach to the prediction of the Vaska's barriers and distances. With the NBO properties, there was a sharp decrease in the MAE at ∼700 dimensions, reaching a minimum at $1.72 \cdot 10^{-2}$ Å ($r^2 = 0.783$) with a 28-dimensional vector, in which the BB and BA terms were 54% of the representation.

The data efficiency of the GP models was also investigated with a training:validation:test data split of 20:40:40 (Table 5). Whereas the accuracy of the resulting models was reasonable for the prediction of the barriers, with MAE = 1.1−1.4 kcal/mol and $r^2 = 0.79 - 0.85$, the performance in the prediction of the distances was rather poor, with MAE = $2.42 \cdot 10^{-2}$ Å and $r^2 < 0.6$, suggesting that the latter regression task must involve further changes in the models before dimensionality reduction can be leveraged efficiently in a small-data training context.

Neural networks

Dimensionality reduction was also explored in the prediction of the Vaska's barriers and distances with NN models. Due to higher computational cost, the recalculation of the GP scatter plots of Figure 7 for the NNs was performed with a smaller resolution (Figure S8). Since the trends observed with both models were similar, the NNs were fed with $v^{I,MD}_{AABBA}$ vectors pruned down to the same optimally reduced dimensions found with the GP (Tables 5 and 6). For the sake of comparability, we used the same NN hyperparameters that yielded the results shown in Tables 1 and 2.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Table 6: Test errors in the prediction of the Vaska's dataset energy barrier ($\Delta E^{\ddagger}$) and H$\cdots$H distance ($d_{H\cdots H}$) with NNs. The inputs passed to the models were $v^{I,MD}_{AABBA}$ vectors defined with different property types (Prop) after being pruned to a reduced dimensionality ($rdim$) based on the accumulated relevance (AR) and MAEs found with the GBM and GP models, respectively. The mean absolute errors (MAEs) are given in kcal/mol for the barriers and Å for the distances.

Target	Input				Average Error[a]		Lowest Error	
	Prop	$rdim$[b]	AA:BB:BA	AR	MAE	r^2	MAE	r^2
$\Delta E^{\ddagger}$	P[c]	46	19:12:15	80%	0.84 ± 0.02	0.926 ± 0.003	0.78	0.935
$\Delta E^{\ddagger}$	P[c]	46	19:12:15	80%	1.37 ± 0.02	0.799 ± 0.007	1.31	0.809
$\Delta E^{\ddagger}$	NBO[d]	115	53:29:33	86%	0.79 ± 0.02	0.912 ± 0.005	0.74	0.931
$\Delta E^{\ddagger}$	NBO[d]	115	53:29:33	86%	1.33 ± 0.01	0.425 ± 0.108	1.29	0.503
$d_{H\cdots H}$	P[c]	100	42:22:36	82%	$1.88\cdot10^{-2} \pm 6\cdot10^{-4}$	0.774 ± 0.013	$1.70\cdot10^{-2}$	0.808
$d_{H\cdots H}$	P[c]	100	42:22:36	82%	$2.77\cdot10^{-2} \pm 9\cdot10^{-4}$	0.473 ± 0.034	$2.53\cdot10^{-2}$	0.572
$d_{H\cdots H}$	NBO[d]	28	13:7:8	52%	$1.94\cdot10^{-2} \pm 4\cdot10^{-4}$	0.744 ± 0.010	$1.84\cdot10^{-2}$	0.781
$d_{H\cdots H}$	NBO[d]	28	13:7:8	52%	$2.50\cdot10^{-2} \pm 2\cdot10^{-4}$	0.563 ± 0.010	$2.46\cdot10^{-2}$	0.577

[a]From ten repetitions with a training:validation:test split of 80:10:10 (black) or 20:40:40 (purple); [b]After pruning the AABBA representation with respect to the GBM relevances; [c]*I.e.* P_A, P_B, and P_{AB} generic property sets; [d]*I.e.* $P_{A,NBO}$, $P_{B,NBO}$, and $P_{AB,NBO}$ NBO property sets. See the Appendix for further details.

In the prediction of the barriers and distances with generic properties, the lowest errors, *i.e.* MAE = 0.78 kcal/mol and $1.70 \cdot 10^{-2}$ Å (Table 6), respectively, were either larger than or similar to those found with the GP (Table 5) and systematic NN (Table 1) models. In the predictions based on NBO properties, the model yielding MAE = 0.74 kcal/mol and $r^2 = 0.931$ for the prediction of the barriers was the most accurate of the series reported in this work. For the distances, the performance level was lower than that of the GP and higher than that of the systematic NNs, with MAE = $1.84\cdot10^{-2}$ (Table 2). In a final experiment, we explored small-data training with the reduced representations using a training:validate:test ratio of 20:40:40, and, in line with the GP results, moderate accuracies were only achieved in the prediction of the energy barriers.

In general, dimensionality could be significantly reduced to increase accuracy with both the GP and NN models, in line with previous observations made for similar ML models leveraging only atom–atom autocorrelations.[45] To some extent, this behavior could be expected in a framework in which the dimensionality of the largest representations is similar to the total number of data points available for training the models. Compared to the systematic

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

models shown in Tables 1 and 2, in which the inputs were manually set, the GBM-engineered AABBA representations achieved higher accuracies in the prediction of both targets with both property types, though often by a small margin. This simplification of the $v^{I,MD}_{AABBA}$ vector did not alter the nature of the properties yielding the lowest MAEs; *i.e.* generic for the prediction of the distances and NBO for the barriers (Tables 5 and 6). Another common trend between the GP and NNs is that the most accurate models using generic properties needed more dimensions for predicting the distances than the barriers, whereas the opposite was true with the NBO properties. These results also showed that, for these regression tasks, accumulating all possible AABBA terms into a large-dimensional vector is not an efficient strategy. Interestingly, in all cases, the bond–bond and bond–atom terms extending the conventional atom–atom autocorrelations accounted for 50-60% of the reduced representation.

Conclusions

In this work, we showed how to extract features from a molecular graph encoding information on both atomic and bond properties. This was implemented by following two distinct approaches: 1) Concatenation of atom–atom, bond–bond, and bond–atom autocorrelation terms, as implemented in the AABBA(I) kernel; and 2) Merging atom and bond properties into vectors that were subsequently autocorrelated into a smaller dimensionality representation, as implemented in the AABBA(II) kernel. The AABBA(I) kernel was implemented in a modular way, allowing for using the AA-AC, BB-AC, $\overline{\text{BB}}$-AC, and BA-AC autocorrelations as stand-alone independent kernels.

The vectors generated by the AABBA kernels were assessed in the prediction of the energy barriers and breaking H$\cdots$H distances of the Vaska's dataset. In a systematic approach gradually adding complexity and dimensionality to the input of NN models, we found that, when used independently, the bond–bond kernels, especially the $\overline{\text{BB}}$-AC, were performing at a level similar to that of the AA-AC baseline. Once all these kernels were combined into

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

AABBA(I), the resulting representation outperformed the AA-AC by a significant margin. High accuracies could be obtained with this kernel using both generic and NBO properties. In general, the influence of the origin, property operator, and maximal depth on the performance of the NNs was either moderate or small. Interestingly, the accuracies achieved with the lower-dimension AABBA(II) kernel were similar (barrier) or higher (distance) than those of AABBA(I), showing the potential benefits of dimensionality reduction in this context.

After maximizing the dimensionality of the representation with the AABBA(I) kernel, GBM models were used to quantify feature relevance in the prediction of the Vaska's barriers and distances. Among the 20 most important features, *ca.* half of them were extracted by the BB-AC, $\overline{\text{BB}}$-AC, and BA-AC components of the kernel, showing the advantage of leveraging bond properties in these regression tasks. Feature relevance also allowed for interpreting the predictions, showing that, whereas the barriers were mostly related to global features encoding electronic structure information, distances were more connected to local metal-centered features encoding bond information. Dimensionality reduction was also explored with GP and NN models in which the dimensions of the AABBA(I) representation were gradually removed according to the feature relevances found by the GBMs. This approach proved efficient, yielding many of the most accurate models reported in this study. In line with the GBM results, $\sim$50% of the reduced autocorrelation vector dimensions were extracted by a kernel component operating on bond properties.

In summary, this work showed that the Moreau-Broto atom–atom autocorrelation kernel on molecular graphs can be extended to include bond–bond and bond–atom terms, increasing the accuracy of the ML models in which the resulting vectors are leveraged as input. For optimal results with the AABBA(I) kernel, we recommend performing dimensionality reduction to simplify the input and achieve higher accuracies. If this feature engineering approach is too involved for the application intended, the AABBA(II) kernel does virtually the same in an implicit and simple manner, achieving similar accuracies. Further, if accuracy is not critical, the stand-alone use of the bond–bond kernels, in particular the $\overline{\text{BB}}$-AC, can

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

also yield satisfactory results at a level close to that of the conventional AA-AC kernel and with the advantage of using a smaller dimensionality representation.

Supporting information

The SI provides further information about the maximal metal-centered depths, computational details of the NN, GBM, and GP models, and additional details about feature relevance and dimensionality reduction.

Data and code

The code of the AABBA graph kernels is openly available at github.com/lmoranglez/AABBA. This URL also provides access to the u-NatQG graphs of the Vaska's complex dataset, the associated AABBA vectors, and the code of all ML models reported in this work. The HyDGL program was used to generate the u-NatQG graphs (github.com/hkneiding/HyDGL).

Author contributions

L.M.-G. implemented the AABBA graph kernels and performed the systematic development of the NN models. J.E.B implemented the feature selection with the GBM algorithm and the associated NN and GP models. H.K. computed the u-NatQG graphs. D.B. designed, supervised, and funded the research project. All authors contributed to scientific discussions, as well as to the writing and revision of the manuscript.

Conflicts of interest

There are no conflicts of interest to declare.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Acknowledgments

L. M.-G. acknowledges Generalitat de Catalunya for a FI-AGAUR predoctoral contract, 2022FI-B2000621. J.E.B acknowledges the Hylleraas Center for hosting his MSc thesis project. H.K. acknowledges the support from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 945371. This work reflects only the author's view and the REA is not responsible for any use that may be made of the information it contains. D.B. acknowledges the support from the FRIPRO (catLEGOS project; number 325003) and SFF (Hylleraas Centre; project number 262695) programs of the Research Council of Norway, and the Norwegian Supercomputing Program (NOTUR; project number NN4654K).

References

(1) Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chem. Sci.* **2018**, *9*, 7069–7077.

(2) Wellendorff, J.; Lundgaard, K. T.; Mogelhoj, A.; Petzold, V.; Landis, D. D.; Norskov, J. K.; Bligaard, T.; Jacobsen, K. W. Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **2012**, *85*, 235149.

(3) Tran, K.; Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution. *Nat. Catal.* **2018**, *1*, 696–703.

(4) Karl, T. M.; Bouayad-Gervais, S.; Hueffel, J. A.; Sperger, T.; Wellig, S.; Kaldas, S. J.; Dabranskaya, U.; Ward, J. S.; Rissanen, K.; Tizzard, G. J. Machine Learning-Guided Development of Trialkylphosphine Ni-(I) Dimers and Applications in Site-Selective Catalysis. *J. Am. Chem. Soc.* **2023**, *145*, 15414–15424.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(5) Nandy, A.; Duan, C.; Goffinet, C.; Kulik, H. J. New Strategies for Direct Methane-to-Methanol Conversion from Active Learning Exploration of 16 Million Catalysts. *JACS Au* **2022**, *2*, 1200–1213.

(6) Gensch, T.; Gomes, G. d. P.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J. Am. Chem. Soc.* **2022**, *144*, 1205–1217.

(7) Pollice, R.; Dos Passos Gomes, G.; Aldeghi, M.; J., H. R.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-Driven Strategies for Accelerated Materials Design. *Acc. Chem. Res.* **2021**, *54*, 849–860.

(8) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

(9) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **2019**, *5*, 83.

(10) Severson, K. A.; Attia, P. M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M. H.; Aykol, M.; Herring, P. K.; Fraggedakis, D. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **2019**, *4*, 383–391.

(11) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2016**, *2*, 16028.

(12) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z. Q.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(13) Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(14) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

(15) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546.

(16) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, *3*, 283–293.

(17) Yang, X.; Wang, Y. F.; Byrne, R.; Schneider, G.; Yang, S. Y. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119*, 10520–10594.

(18) Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* **2015**, *20*, 318–331.

(19) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559.

(20) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep learning for computational chemistry. *J. Comput. Chem.* **2017**, *38*, 1291–1307.

(21) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **2018**, *4*, eaap7885.

(22) Simm, G. N.; Reiher, M. Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes. *J. Chem. Theory Comput.* **2018**, *14*, 5238–5248.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(23) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121*, 10073–10141.

(24) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89–96.

(25) Zheng, C.; Chen, C.; Chen, Y. M.; Ong, S. P. Random Forest Models for Accurate Identification of Coordination Environments from X-Ray Absorption Near-Edge Structure. *Patterns* **2020**, *1*, 100013.

(26) Sheridan, R. P.; Wang, w. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353–2360.

(27) Li, L.; Zhao, Y.; Yu, H.; Wang, Z.; Zhao, Y.; Jiang, M. An XGBoost Algorithm Based on Molecular Structure and Molecular Specificity Parameters for Predicting Gas Adsorption. *Langmuir* **2023**, *39*, 6756–6766.

(28) Chen, S.; Nielson, T.; Zalit, E.; Skjelstad, B. B.; Borough, B.; Hirschi, W. J.; Yu, S.; Balcells, D.; Ess, D. H. Automated Construction and Optimization Combined with Machine Learning to Generate Pt(II) Methane C-H Activation Transition States. *Top. Catal.* **2022**, *65*, 312–324.

(29) Mayr, A.; Klambauer, G.; Unterthiner, S., T Hochreiter DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 80.

(30) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(31) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

(32) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **2018**, *148*, 241718.

(33) Huang, B.; von Lilienfeld, O. A. Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *J. Chem. Phys.* **2016**, *145*, 161102.

(34) De, S.; Bartok, A. P.; Csanyi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

(35) Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **2018**, *57*, 13973–13986.

(36) Nandy, A.; Duan, C. R.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. *Chem. Rev.* **2021**, *121*, 9927–10000.

(37) Balcells, D.; Skjelstad, B. B. The tmQM Dataset - Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135–6146.

(38) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(39) Krenn, F., M Häse; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(40) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **2021**, *12*, 2198–2208.

(41) Kneiding, H.; Lukin, R.; Lang, L.; Reine, S.; Pedersen, T. B.; de Bin, R.; Balcells, D. Deep Learning Metal Complex Properties with Natural Quantum Graphs. *Digital Discovery* **2023**, *2*, 618–633.

(42) Rupp, M.; Schneider, G. Graph Kernels for Molecular Similarity. *Mol. Inf.* **2010**, *29*, 266–273.

(43) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors. Vol. 11.*; Weinheim, Germany: Wiley-VCH., 2008.

(44) Moreau, G.; Broto, P. Autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359–360.

(45) Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939–8954.

(46) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513–524.

(47) Glendening, E. D.; Landis, C. R.; Weinhold, F. NBO 7.0: New vistas in localized and delocalized chemical bonding theory. *J. Comput. Chem.* **2019**, *40*, 2234–2241.

(48) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proc. Mach. Learn. Res.* **2017**, *70*, 1263–1272.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

(49) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757–1772.

(50) Friederich, P.; dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584–4601.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Appendix

Atom–Atom Autocorrelation

The original atom–atom autocorrelation (AA-AC) of Moreau and Broto[44] (Figure 2) transforms any connected molecular graph $\mathcal{G}$ into a fixed-length vector, v_{AA}, regardless of the size of $\mathcal{G}$. The calculation of v_{AA} is based on the autocorrelation function

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{i=1}^{N_{\mathcal{G}}} \sum_{j=1}^{N_{\mathcal{G}}} p_i p_j \delta_{d,d_{i,j}} \tag{2}$$

where $N_{\mathcal{G}}$ is the number of atomic nodes in the molecular graph, p is an atomic property (*e.g.* the atomic number), i and j are atomic indices, p_i and p_j are the correlated properties of atoms i and j, and δ is the Kronecker delta; *i.e.* $\delta_{d,d_{i,j}} = 1$ for $d = d_{i,j}$ and 0 for $d \neq d_{i,j}$, where d, the depth, is the distance in the number of edges (*i.e.* chemical bonds) along the shortest path connecting the atomic nodes i and j. The f_{AC} function is permutation invariant relative to the (i, j) indices of $\mathcal{G}$.

The visual intuition behind Equation 2 is the use of the skeletal formula of a TMC as a computational graph (Figure 2) in which the properties of the atoms are correlated by multiplication to those of the neighborhood at a given depth, adding the resulting values to obtain the components of the autocorrelation v_{AA} vector. Since, in general, the dimensionality of this vector is smaller than that of the associated graph, the $\mathcal{G} \Rightarrow v_{AA}$ transformation can be seen as a data compression operation yielding a molecular fingerprint.

The v_{AA} vector is generated by collecting the f_{AC} values at different depths (Figure 2), as shown in Equation 3

$$v_{AA}(D) = (f_{AC}(d=0), f_{AC}(d=1), ..., f_{AC}(d=D)) \tag{3}$$

where d is expanded from $d = 0$ to the maximum depth of the representation, D, in $+1$ increments; *i.e.* $d \in \{0, 1, 2..., D\}$.

The v_{AA} vector is further extended by expanding the property p to a set of K atomic properties, P_A; *i.e.*

$$P_A = \{P_{A,1}, P_{A,2}, P_{A,3}..., P_{A,K}\} \tag{4}$$

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

from which, considering the depth, results

$$v_{AA}(P, D) = \\ (f_{AC}(P_{A,1}, d = 0), ..., f_{AC}(P_{A,1}, d = D), ..., f_{AC}(P_{A,K}, d = 0), ..., f_{AC}(P_{A,K}, d = D)) \tag{5}$$

with an overall dimensionality of

$$dim(v_{AA}) = (D + 1) \times K \tag{6}$$

For example, the use of the atomic number (Z) and covalent radius (R) as properties (*vide infra*) for a maximum depth of 3 yields the following eight-dimensional autocorrelation vector:

$$v_{AA} = (Z_0, Z_1, Z_2, Z_3, R_0, R_1, R_2, R_3) \tag{7}$$

In addition to the depth and the atomic properties, the autocorrelation algorithm depends on two more variables; namely 1) the definition of the $d = 0$ origin, and 2) the arithmetic operator applied to the properties.

For mononuclear TMCs, the metal atom is a natural and unambiguous choice for setting the depth origin (Figure 3) from which metal-centered (MC) autocorrelations can be computed with this equation:

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{j=1}^{N_{\mathcal{G}}} p_M p_j \delta_{d,d_{M,j}} \tag{8}$$

where M is the metal center index. The other possibility is to do a full (F) autocorrelation in which all nodes are recursively used as the $d = 0$ origin once (*i.e.* Equation 2). Whereas the full AA-AC can compress more information into the v_{AA} vector, the metal-centered flavor can express electronic and steric properties over the {α, β, γ, ...} positions around the metal center, which are equivalent to $d = 0, 1, 2, ...$, in a way that organometallic and inorganic chemists can relate intuitively to proximal and distal effects.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Regarding the arithmetic operator, and besides the product autocorrelation ($\odot$), which is the one most commonly used, division, summation, and subtraction,

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{i=1}^{N_{\mathcal{G}}} \sum_{j=1}^{N_{\mathcal{G}}} \frac{p_i}{p_j} \delta_{d,d_{i,j}} \tag{9}$$

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{i=1}^{N_{\mathcal{G}}} \sum_{j=1}^{N_{\mathcal{G}}} (p_i + p_j) \delta_{d,d_{i,j}} \tag{10}$$

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{i=1}^{N_{\mathcal{G}}} \sum_{j=1}^{N_{\mathcal{G}}} (p_i - p_j) \delta_{d,d_{i,j}}, \tag{11}$$

can also be used and referred to as ratiometric ($\oslash$; Equation 9), summetric ($\oplus$; Equation 10), and deltametric ($\ominus$; Equation 11) autocorrelations.

For TMCs, another possibility is to compute autocorrelations with distinct scopes reflecting the coordination geometry.[45] For example, for a trigonal bipyramid TMC, it is possible to define separate terms for the axial and equatorial ligands. This option was not considered in the present study since we were interested in generalizing the AABBA graph kernels over datasets containing a wide range of different coordination geometries. For example, the tmQMg dataset[41] contains thousands of linear, bent, trigonal planar, tetrahedral, square planar, trigonal bipyramid, square pyramid, and octahedral complexes.

Bond–Bond Autocorrelation

Adding to the AA-AC term, and inspired by the donor-acceptor interactions between bond orbitals in NBO analysis (Figure 1), we developed the bond–bond autocorrelation concept (BB-AC; Figure 2). In the full BB-AC implementation, bond properties are autocorrelated with the same f_{AC} function used to compute AA-AC, considering all bonds as the depth origin once. For this purpose, Equation 2 is reinterpreted as if the chemical bonds were the graph nodes; *i.e.* $N_{\mathcal{G}}$ is the number of bonds in the molecular graph, p is a bond property, i and j are bond indices, p_i and p_j are the correlated properties of bonds i and j, and, in the Kronecker delta $\delta_{d,d_{i,j}}$, d is the distance in the number of atoms along the shortest path connecting bonds i and j.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

At the computational graph level (Figure 2), the visual interpretation of BB-AC is analogous to that of AA-AC, now feeding the product operator with bond properties instead of atom properties. As for v_{AA}, the BB-AC autocorrelation vector, v_{BB}, is composed with Equation 5 after collecting the values of its components at different depths for different properties, which can be calculated using any of the four arithmetic operators (equations 2 and 9-11). The set of properties is

$$P_B = (P_{B,1}, P_{B,2}, P_{B,3}..., P_{B,L}) \tag{12}$$

which contains L bond properties and, in general, $L < K$ because there are more properties available to describe the atoms than the bonds. The resulting dimensionality is thus

$$dim(v_{AC}^{BB}) = (D + 1) \times L \tag{13}$$

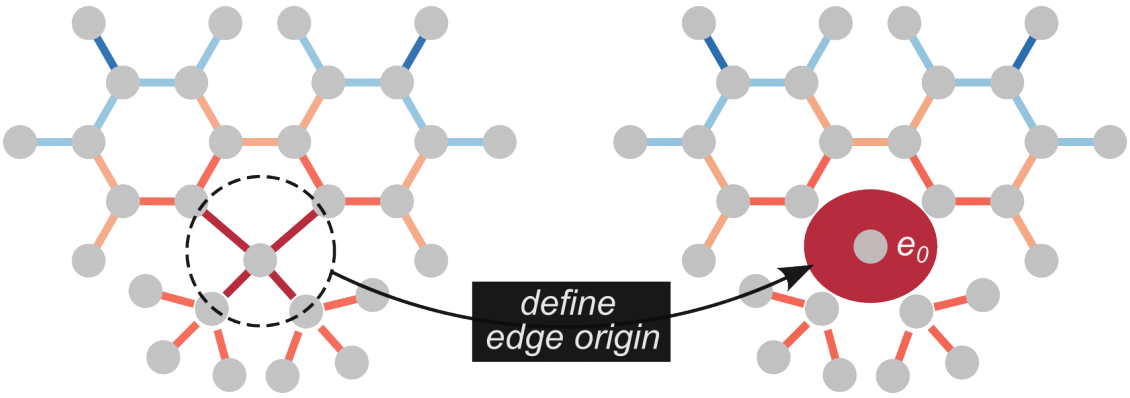

Figure 8: Definition of the edge origin e_0 (red ellipse) in the metal-centered **BB-AC** autocorrelation. The dotted-line black circle comprises the metal center and the bonds connected to it.

The metal-centered BB-AC requires redefining the depth origin (Figure 3). Whereas in AA-AC the origin is trivial and unique for any mononuclear TMC, in BB-AC the origin involves several bonds connecting the metal center to the ligands. Figure 8 illustrates how we implemented the BB-AC depth origin; the index of the metal center is used to identify the bonds involving it, which, as a whole, form the edge origin, e_0, at $d = 0$; *i.e.* e_0 is the following set of metal-ligand bond edges:

$$e_0 = \{e_{0,1}, e_{0,2}, ..., e_{0,CN}\} \tag{14}$$

where CN is the coordination number of the metal center.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The properties of e_0, p_{e_0}, are calculated by either averaging the properties of the edge set:

$$\overline{p}_{e_0}(CN, p) = \frac{\sum_{i=1}^{CN} p_{e_{0,i}}}{CN} \tag{15}$$

or summing them up:

$$p_{e_0}(CN, p) = \sum_{i=1}^{CN} p_{e_{0,i}} \tag{16}$$

Once the edge origin and its properties are defined, the metal-centered BB-AC autocorrelations are calculated with this function:

$$f_{AC}(N_{\mathcal{G}}, p, d) = \sum_{j=1}^{N_{\mathcal{G}}} \mathcal{P}_{e_0} p_j \delta_{d,d_{e_0,j}} \tag{17}$$

and the resulting v_{BB} vector, which collects all property and depth dimensions as in Equation 5, is labeled either $\overline{\text{BB}}$-AC or BB-AC, depending on whether $\mathcal{P}_{e_0}$ is equal to $\overline{p}_{e_0}$ or p_{e_0}, respectively.

Bond–Atom Autocorrelation

Further adding to the AA-AC and BB-AC terms, we also implemented the bond–atom autocorrelations (BA-AC; Figure 2), which were inspired by both NBO analysis and the coupling between embedded atom and bond properties in message-passing graph neural networks (Figure 1). The full BA-AC is implemented with this equation:

$$f_{AC}(N_{\mathcal{G},\mathcal{V}}, N_{\mathcal{G},\mathcal{E}}, p, d) = \sum_{i=1}^{N_{\mathcal{G},\mathcal{V}}} \sum_{j=1}^{N_{\mathcal{G},\mathcal{E}}} p_i p_j \delta_{d,d_{i,j}} \tag{18}$$

where $N_{\mathcal{G},\mathcal{V}}$ and $N_{\mathcal{G},\mathcal{E}}$ are the number of atomic nodes and bond edges in the molecular graph, respectively, p_i and p_j are the correlated properties of atom i and bond j, and, in the Kronecker delta $\delta_{d,d_{i,j}}$, d is the distance in number of atoms between node i and bond j. The computational graph underlying Equation 18 is shown in Figure 2.

The metal-centered BA-AC is computed with this equation:

$$f_{AC}(N_{\mathcal{G},\mathcal{E}}, p, d) = \sum_{j=1}^{N_{\mathcal{G},\mathcal{E}}} p_M p_j \delta_{d,d_{M,j}} \tag{19}$$

where p_M is an atomic property of the metal center. As for AA-AC and BB-AC, the BA-AC autocorrelation vector, v_{BA}, is composed using Equation 5 to gather all depth and property dimensions.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The product between the atomic and bond properties, which belong to sets of different dimensionality (Equations 4 and 12), was implemented as follows:

$$p_i p_j = \sum_{l=1}^{L} p_i p_{j,l} \tag{20}$$

where, for p_j, L is the dimensionality of the bond property set and, for p_i, i is either any graph node (full BA-AC) or the metal (metal-centered BA-AC). Thus, in both cases, the final dimensionality of the representation is

$$dim(v_{BA}) = (D+1) \times K \tag{21}$$

where K is the dimensionality of the atomic property set.

The mixing of atomic and bond properties in Equation 18 may suggest that the term crosscorrelation could be more appropriate than autocorrelation for referring to BA-AC. However, since both terms have additional and different meanings in the field of signal processing, we decided to keep the autocorrelation term originally proposed by Moreau and Broto for molecular graphs.[44]

Atomic and bond properties

The v_{AA} vector can be derived from a set of features including atomic properties (P_A) that can be generic (*e.g.* extracted from the periodic table). In this work, we used this popular P_A set for TMCs:

$$P_A = \{Z, I, V, R, \chi\} \tag{22}$$

From a chemical perspective, the most relevant properties are the atomic number (Z), the covalent radius (R), and the electronegativity (χ). The properties I and V are relevant from both a chemical and a graph theory perspective: V is the atomic valence, which is equal to the node degree (*i.e.* number of neighbors connected to a node), and I, which is either 0 or 1, indicates the absence or presence, respectively, of a node at any given depth in the graph paths walked by the autocorrelation algorithm (Equation 2).

With this simple set of properties, the AA-AC autocorrelations already provide rich information about the systems they encode, including chemical composition and environment, through Z and V, and steric bulk, through R. Further, the variation of I over $d = 0, 1, ..., D$ reflects the shape of the TMC (*e.g.* linear versus branched), and, by changing the arithmetic operator, additional information can be included in the v_{AA} vector; *e.g.*, bond polarization can be encoded by applying the subtraction operator (Equation 11) to χ (*i.e.* deltametric electronegativity).

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

For the v_{BB} vector, we used this set of bond properties:

$$P_B = \{BO, I, BD\} \tag{23}$$

to compute the BB-AC autocorrelations, where BO is the bond order, I is the identity, which has the same meaning as in P_A, and BD is the bond distance in Å, which gives further geometric information in addition to that provided by R and I in P_A.

Table 7: NBO properties included in the atomic ($P_{A,NBO}$) and bond ($P_{B,NBO}$) property sets.[a]

	$P_{A,NBO}$		$P_{B,NBO}$
Z	Atomic number	BD	Bond distance (Å)
q_{Nat}	Natural charge (e)	BO_{Nat}	Natural Wiberg bond order
V_{Nat}	Natural valence index	N_{BN}	# bonding NBOs
N_s	# s electrons in nat. config.	BN_E	E of highest-lying BN (Ha)
N_p	# p electrons in nat. config.	$BN_{\Delta E}$	Lowest/highest-lying BN E gap (Ha)
N_d	# d electrons in nat. config.	BN_{Occ}	Electron occupancy of highest-E BN
N_{LP}	# Lone pairs	BN_s	s-character of highest-E BN (%)
LP_E	E of highest-lying LP (Ha)	BN_p	p-character of highest-E BN (%)
$LP_{\Delta E}$	Lowest/highest-lying LP E gap (Ha)	BN_d	d-character of highest-E BN (%)
LP_{Occ}	Electron occupancy of highest-E LP	N_{BN^*}	# non- & anti-bonding NBOs
LP_s	s-character of highest-E LP (%)	BN^*_E	E of lowest-lying BN* (Ha)
LP_p	p-character of highest-E LP (%)	$BN^*_{\Delta E}$	Lowest/highest-lying BN* E gap (Ha)
LP_d	d-character of highest-E LP (%)	BN^*_{Occ}	Electron occupancy of lowest-E BN*
N_{LV}	# Lone vacancies	BN^*_s	s-character of lowest-E BN* (%)
LV_E	E of lowest-lying LV (Ha)	BN^*_p	p-character of lowest-E BN* (%)
$LV_{\Delta E}$	Lowest/highest-lying LV E gap LV (Ha)	BN^*_d	d-character of lowest-E BN* (%)
LV_{Occ}	Electron occupancy of lowest-E LV		
LV_s	s-character of lowest-E LV (%)		
LV_p	p-character of lowest-E LV (%)		
LV_d	d-character of lowest-E L (%)		

[a]Abbreviations: # = Number of; E = Energy; Nat. = Natural; LP = Lone Pair; LV = Lone Vacancy; Config. = Configuration; BO = Bond Order; NBOs = Natural Bond Orbitals; BN = Bonding NBO; BN* = Non- and anti-bonding NBOs.

Except for the bond distance, the properties included in the P_A and P_B sets are generic, thus having a limited capacity in distinguishing different chemical environments; *e.g.* the C atom of any R–CH_2–R' fragment is described with the same P_A values regardless of the nature of R and R'. This limitation can be tackled by using specific electronic structure properties from inexpensive quantum mechanical calculations. In a recent study, we showed that the leverage of this electronic structure information in graph neural networks boosts the prediction accuracy of the resulting models to an extent larger than that provided by geometric information.[41] In the present work, we investigated the use of NBO data in the computation of the autocorrelation vectors.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The set of NBO atomic properties, $P_{A,NBO}$, includes the atomic number, the natural charge and valence index, the number of lone pairs (LP) and vacancies (LV), and the electron occupancies and symmetries of the highest- and lowest-energy LP and LV orbitals, respectively. The set of NBO bond properties, $P_{B,NBO}$, includes the same information for the bonding and antibonding valence orbitals, as well as the natural bond order and the bond distance. Table 7 provides a systematic list of all NBO data included in $P_{A,NBO}$ and $P_{B,NBO}$.

Whole-graph properties

We defined a set of whole-graph properties; *i.e.*

$$P_{\mathcal{G}} = \{q, M, N_{At}, N_e\}, \tag{24}$$

which contains the charge of the metal complex (q), its molecular mass (M), and the total number of atoms (N_{At}) and electrons (N_e). These properties were appended to the end of the AABBA autocorrelation vector (*vide infra*) in the ML models:

$$v^{\mathcal{G}}_{AABBA} = v_{AABBA} \oplus P_{\mathcal{G}} \tag{25}$$

where $\oplus$ denotes the vector concatenation operation and v_{AABBA} is either v^{I}_{AABBA} or v^{II}_{AABBA}, as described in the next section.

Atom–Atom Bond–Bond Bond–Atom Autocorrelations

With the aim of developing a molecular graph-to-vector transformation in which both atom and bond properties are autocorrelated separately and jointly, we developed an atom–atom bond–bond bond–atom AABBA graph kernel yielding v_{AABBA} autocorrelation vectors through two distinct implementations that can be regarded as being either explicit or implicit.

In the explicit implementation of the graph kernel, AABBA(I), the resulting vector representation, v^{I}_{AABBA}, was composed by simply joining the AA-AC, $\overline{\mathrm{BB}}$-AC, and BA-AC autocorrelations as follows:

$$v^{I}_{AABBA} = v_{AA} \oplus v_{\overline{BB}} \oplus v_{BA} \tag{26}$$

which has dimensionality

$$dim(v^{I}_{AABBA}) = (D+1) \times (2K+L) \tag{27}$$

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

where D = maximum depth, $K = dim(P_A)$, and $L = dim(P_B)$; the NBO $P_{A,NBO}$ and $P_{B,NBO}$ property sets can be also correlated, expanding the dimensionality of the resulting representations.

In the implicit implementation, AABBA(II), the v^{II}_{AABBA} vector was computed with the bond–bond autocorrelation function, as defined in Equation 17, applied to property sets that describe both the bond and the atoms associated to it (P_{AB}); in particular, for any i–j bond edge connecting the atomic nodes i and j, we considered these three sets based on generic properties:

$$\begin{aligned} P_{AB,1} &= \{Z_i, Z_j, V_i, V_j, \chi_i, \chi_j, BD, BO, I\}; M = 9 \\ P_{AB,2} &= \{Z_i, Z_j, V_i, V_j, \chi_i - \chi_j, BD, BO, I\}; M = 8 \\ P_{AB,3} &= \{Z_i, Z_j, V_i, V_j, \chi_i - \chi_j, R_i, R_j, BO, I\}; M = 9 \end{aligned} \tag{28}$$

In $P_{AB,1}$, each bond is described by its distance and order, whereas the associated atoms are described by their atomic number, valence, and electronegativity. In $P_{AB,2}$, the latter is replaced by the $\chi_i - \chi_j$ difference, which accounts for the polarization of the i–j bond. Lastly, in $P_{AB,3}$, the bond distance is replaced by the covalent radii of the atoms to yield a geometry-agnostic representation. We also defined two additional P_{AB} sets based on NBO properties:

$$\begin{aligned} P_{AB,4} = \{&q_{Nat,i}, q_{Nat,j}, V_{Nat,i}, V_{Nat,j}, N_{s,i}, N_{s,j}, N_{p,i}, N_{p,j}, N_{d,i}, N_{d,j}, N_{LP,i}, N_{LP,j}, N_{LV,i}, N_{LV,j}, \\ &BD, BO_{Nat}, N_{BN}, BN_s, BN_p, BN_d, N_{BN^*}, BN^*_s, BN^*_p, BN^*_d, I\}; M = 25 \end{aligned} \tag{29}$$

$$\begin{aligned} P_{AB,5} = \{&q_{Nat,i}, q_{Nat,j}, V_{Nat,i}, V_{Nat,j}, N_{LP,i}, N_{LP,j}, LP_{E,i}, LP_{E,j}, LP_{\Delta E,i}, LP_{\Delta E,j}, \\ &N_{LV,i}, N_{LV,j}, LV_{E,i}, LV_{E,j}, LV_{\Delta E,i}, LV_{\Delta E,j}, BD, BO_{Nat}, N_{BN}, BN_E, BN_{\Delta E}, \\ &N_{BN^*}, BN^*_E, BN^*_{\Delta E}, I\}; M = 25 \end{aligned} \tag{30}$$

where $P_{AB,4}$ is rich in orbital symmetry information whereas $P_{AB,5}$ is rich in orbital energy information.

The resulting autocorrelations were labeled AABBA(II)$_n$, where n is the index of the $P_{AB,n}$ property set used in their calculation. The dimensionality of the associated vectors is

$$dim(v^{II}_{AABBA}) = (D+1) \times M \tag{31}$$

where M is the number of properties included in the P_{AB} sets, as shown in Equations 28, 29 and 30. Both the AABBA(I) and AABBA(II)$_n$ kernels are available in the full and metal-centered flavors.

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Maximal dimensionality autocorrelation vectors

With the aim of selecting and interpreting features with GBM models, we extended the autocorrelation vectors to maximal dimensionality (MD). For both the generic and NBO properties, separately, we used the AABBA(I) graph kernel to compute the $v^{I,MD}_{AABBA}$ autocorrelation vectors, with the concatenation operation defined in Equation 26, and including both the BB-AC and $\overline{\text{BB}}$-AC autocorrelations in the bond–bond term. Further, all terms were expanded in both full and metal-centered fashions, and using, in this order, the product, subtraction, division, and summation operators; for example, for the atom–atom autocorrelation:

$$v_{AA} \in v^{I,MD}_{AABBA} = (v^{\odot}_{AA} \oplus v^{\ominus}_{AA} \oplus v^{\oslash}_{AA} \oplus v^{\oplus}_{AA})_{full} \oplus (v^{\odot}_{AA} \oplus v^{\ominus}_{AA} \oplus v^{\oslash}_{AA} \oplus v^{\oplus}_{AA})_{MC} \tag{32}$$

in which each vector component was expanded from depth zero to six (Figure S1).

https://doi.org/10.26434/chemrxiv-2023-5wbkr **ORCID:** https://orcid.org/0000-0002-3389-0543 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0