

RESEARCH ARTICLE | AUGUST 15 2024

## DASH properties: Estimating atomic and molecular properties from a dynamic attention-based substructure hierarchy

Special Collection: [Machine Learning for Biomolecular Modeling](#)

Marc T. Lehner ; Paul Katzberger ; Niels Maeder ; Gregory A. Landrum ; Sereina Riniker



*J. Chem. Phys.* 161, 074103 (2024)

<https://doi.org/10.1063/5.0218154>



View  
Online



Export  
Citation

### Articles You May Be Interested In

Virtual interface substructure synthesis method for normal mode analysis of super-large molecular complexes at atomic resolution

*J. Chem. Phys.* (October 2011)

Quantum dissipative systems beyond the standard harmonic model: Features of linear absorption and dynamics

*J. Chem. Phys.* (October 2019)

*Ab initio* dispersion potentials based on physics-based functional forms with machine learning

*J. Chem. Phys.* (May 2024)

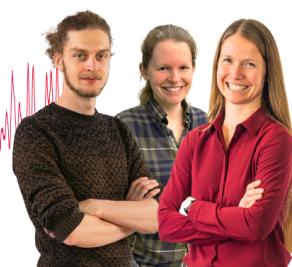
### Webinar From Noise to Knowledge

May 13th – Register now



Zurich  
Instruments

Universität  
Konstanz



# DASH properties: Estimating atomic and molecular properties from a dynamic attention-based substructure hierarchy

Cite as: J. Chem. Phys. 161, 074103 (2024); doi: 10.1063/5.0218154

Submitted: 9 May 2024 • Accepted: 1 August 2024 •

Published Online: 15 August 2024



View Online



Export Citation



CrossMark

Marc T. Lehner, Paul Katzberger, Niels Maeder, Gregory A. Landrum, and Sereina Riniker<sup>a)</sup>

## AFFILIATIONS

Department of Chemistry and Applied Biosciences, ETH Zürich, Vladimir-Prelog-Weg 2, 8093 Zürich, Switzerland

**Note:** This paper is part of the JCP Special Topic on Machine Learning for Biomolecular Modeling.

<sup>a)</sup>Author to whom correspondence should be addressed: [sriniker@ethz.ch](mailto:sriniker@ethz.ch)

## ABSTRACT

Recently, we presented a method to assign atomic partial charges based on the DASH (dynamic attention-based substructure hierarchy) tree with high efficiency and quantum mechanical (QM)-like accuracy. In addition, the approach can be considered “rule based”—where the rules are derived from the attention values of a graph neural network—and thus, each assignment is fully explainable by visualizing the underlying molecular substructures. In this work, we demonstrate that these hierarchically sorted substructures capture the key features of the local environment of an atom and allow us to predict different atomic properties with high accuracy without building a new DASH tree for each property. The fast prediction of atomic properties in molecules with the DASH tree can, for example, be used as an efficient way to generate feature vectors for machine learning without the need for expensive QM calculations. The final DASH tree with the different atomic properties as well as the complete dataset with wave functions is made freely available.

© 2024 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0218154>

## I. INTRODUCTION

The properties of atoms in molecules are often used to describe or estimate the properties of the corresponding molecule. For instance, atomic properties, such as partial charges, are valuable for fixed-charge force fields used in molecular dynamics (MD) simulations or for machine learning (ML) of molecular properties, such as binding affinity or aqueous solubility, where atomic properties can serve as input features. However, in many of these examples, either simple tabulated values are employed or the accuracy of the atomic property is dictated by the computational cost of the quantum-mechanical (QM) calculation required to obtain it. In case of atomic partial charges, this issue can be circumvented by training ML models to predict the partial charges of new molecules.<sup>1–5</sup> Although this approach provides the desired computational efficiency while retaining the underlying QM accuracy, it often lacks explainability and is dependent on the chosen software.

Recently, we presented the dynamic attention-based substructure hierarchy (DASH),<sup>6</sup> which is a tree structure built to assign partial charges to atoms based on 2D topological substructures. The

substructures are expanded iteratively from the initial atom guided by attention values that represent the importance of neighbors and linearize the search, making the method comparable in accuracy to state-of-the-art ML models but even faster. Importantly, the DASH tree has the added benefit of full explainability and error estimates of the assigned charges based on the matched substructure. The DASH partial charges were found to have small errors compared to the minimal basis iterative stockholder (MBIS)<sup>7</sup> reference values, even when tested on the external test set VEHICLe (virtual exploratory heterocyclic library),<sup>8</sup> which demonstrated the generalizability of the DASH approach to unseen molecules. In many cases, the inherent 2D nature of the predictions with the DASH tree can be an advantage. In the case of partial charges for fixed-charge force fields, they need to be conformationally averaged, which is in other force fields handled by performing multiple QM calculations,<sup>9</sup> which adds to the computational cost.

With the DASH tree at hand, the following question arose: how much of the electronic neighborhood of an atom is captured by the DASH tree and how transferable these learned substructures are? Besides the partial charge, other atomic properties (i.e.,

atomic polarizability and dispersion) depend on the local electronic environment of an atom. In this work, we demonstrate that such additional atomic properties can be assigned with the same DASH tree (built for MBIS partial charges), without the need to re-train the associated graph neural network and/or re-build the tree. These properties include other partial charge models (Mulliken,<sup>10</sup> AM1-BCC,<sup>11</sup> and RESP<sup>12</sup>), dispersion ( $C_6$ ,<sup>13</sup> atomic polarizability, and atomic parameters to describe the electro- and nucleophilicity (i.e., Fukui<sup>14</sup> and the dual descriptor<sup>15</sup>). Many of these properties were evaluated directly on the wave functions calculated for the DASH dataset,<sup>6,16</sup> while for others, semi-empirical methods were used. In addition, we were able to speed up the prediction of these properties with the DASH tree by orders of magnitude with improvements in the tree implementation and algorithm, making the DASH approach an exceptional source of features for ML approaches on organic molecules.

## II. METHODS

We use the dataset from the previous work by Lehner *et al.*<sup>6,16</sup> consisting of 348 935 diverse molecules (determined by Morgan fingerprints<sup>29</sup> with radius two). For each molecule, the wave functions of up to three conformers using a TPSSh/def-TZVP functional and basis set (obtained with Psi4<sup>17</sup>) were available, allowing us to efficiently calculate additional properties. These were chosen based on their expected correlation with the local electronic environment of the atom and the availability of computational methods for generating them. A schematic depiction of the two-step workflow is shown in Fig. 1. First, the properties were calculated for all molecules in the DASH dataset<sup>16</sup> and the external test set (VEHICLE<sup>8</sup>). Second, the DASH tree previously built for the assignment of MBIS partial charges was populated with these additional properties. To

accommodate the new data, the implementation of the DASH tree was optimized for memory efficiency and runtime. The improvements in runtime over the previous implementation are discussed in Sec. S1 and shown in Fig. S1 of the [supplementary material](#).

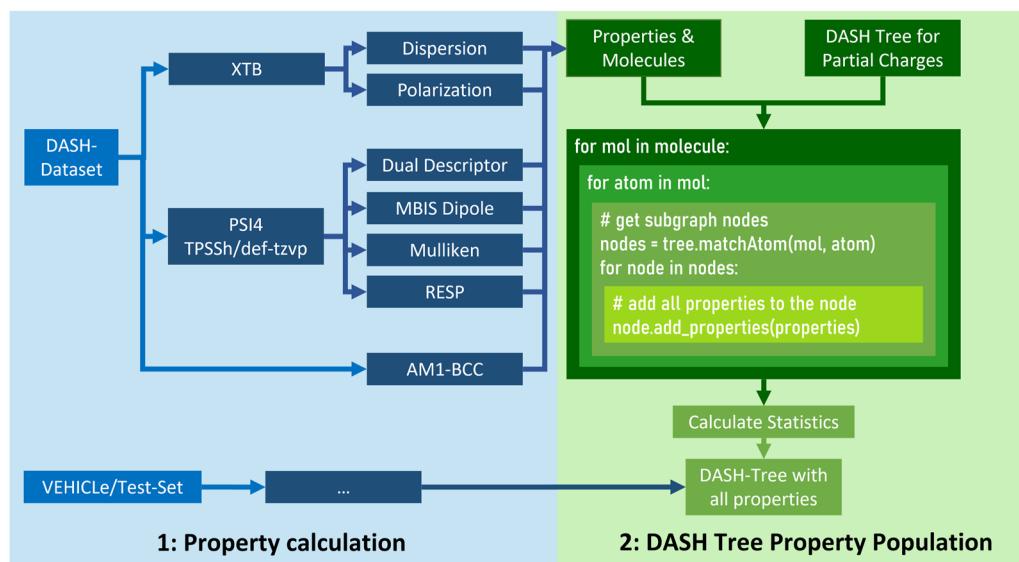
### A. Test sets

In Ref. 6, the dataset was randomly split into a training set and validation set. We made use of the same split; however, since no classical training was performed here (i.e., the validation set was not used for any meta- or hyperparameter tuning), we consider the validation set from Ref. 6 as a *de facto* test set for this work. In addition, the VEHICLE set<sup>8</sup> with 24 657 heterocycles was used again as an external test set to validate all results. The 56 molecules that are in both the DASH dataset and the VEHICLE set were removed from the latter. The results with the DASH test set are shown in the main text, while the results for the VEHICLE test set are shown in the [supplementary material](#) (Figs. S5–S7).

### B. Calculation of the atomic properties

#### 1. Properties at DFT level of theory

The previously calculated wave functions could be used directly for the calculation of additional properties. The function *oeprop* in Psi4 was used to compute Mulliken charges,<sup>10</sup> and the Psi4 RESP package was used to calculate RESP (restrained electrostatic potential) partial charges.<sup>12</sup> We previously<sup>6</sup> showed that the differences between the commonly used B3LYP functional and TPSSh functional used for the DASH tree are small for the calculation of RESP charges. From the calculation of the MBIS<sup>7</sup> partial charge, the dipole moment of the atomic charge distribution could be recovered. The projection of the atomic dipole vectors to the bond vectors of an



**FIG. 1.** Schematic depiction of the workflow to populate a pre-built DASH tree<sup>6</sup> (built to assign MBIS<sup>7</sup> partial charges) with additional computed properties. In the first step (blue, left), the available wave functions were used to calculate new properties. In the second step (green, right), the nodes of the DASH tree were populated with the new properties.

atom was used for the calculation of the molecular dipoles according to Eq. (3) as discussed below. The dual descriptor<sup>15</sup> was calculated from the wave function using the Psi4 function cubeProp with the default grid spacing of 0.2 bohrs.

## 2. Properties at semi-empirical level of theory

To calculate the dispersion and polarizability of each atom, we employed DFTD4.<sup>13</sup>

In addition, AM1-BCC<sup>11</sup> partial charges were calculated due to their usage in many force fields.<sup>9,18</sup> These partial charges were calculated with the OpenFF toolkit (version 0.10.0)<sup>19</sup> using the Amber toolkit (version 22.0).<sup>20</sup>

## C. Populating the DASH tree with new atomic properties

New properties were assigned to the nodes in the existing DASH tree in the same way as the MBIS partial charges were previously assigned. This means that the new atomic properties were assigned to the same nodes as the corresponding MBIS partial charge of the atom. After all molecules were matched against the tree, the median and variance values of each property in each node were calculated. By reusing the DASH tree structure built for the MBIS partial charges, we make the assumption—and test it—that this DASH tree captures the important aspects of an atom's electronic neighborhood, which are applicable to other atomic properties. This allows the prediction of different properties on the basis of the same substructures.

Note that this population process does not include any fitting. The DASH tree is a classification model that groups atom environments in a hierarchical structure. Each node's property is simply the median of all properties of all atoms matching this node. Due to this classification based on 2D molecular substructures, DASH cannot accurately predict properties that have large 3D conformational dependencies or involve strong delocalization, which requires a treatment of the full molecular orbitals.

## D. Calculation of molecular properties

While the atomic properties are straightforward to assign using the atom-centered substructures in the DASH tree, assigning molecular properties requires us to be able to decompose the property value  $p$  into atomic contributions, which can then be combined using a function  $f$ ,

$$p = f(a_{1,p}, a_{2,p}, \dots, a_{N,p}). \quad (1)$$

The function  $f$  can have many forms, but ideally, it would be a simple equation in order to allow for fast property assignment. Here, we focused on two molecular properties for which such a function  $f$  exists. For both, the DASH tree was used to predict the atomic contributions, and then, the molecular property was calculated with the corresponding function  $f$ .

### 1. Molecular polarizability

The molecular polarizability  $\chi_{mol}$  can simply be calculated from the sum of atomic polarizabilities  $\chi_i$ ,<sup>13,21</sup>

$$\chi_{mol} = \sum_i^N \chi_i. \quad (2)$$

While the molecular polarizability is experimentally measurable, the agreement between QM calculations and experimental values is not very high;<sup>22</sup> thus, we compare the values estimated from the DASH tree with the calculated reference values.

### 2. Molecular dipole

Another experimentally measurable property with a simple combination function is the molecular dipole,

$$d_{mol} = \sum_i^N q_i \vec{r}_i + \vec{d}_i \approx \sum_i^N q_i \vec{r}_i, \quad (3)$$

where  $q_i$  is the partial charge of atom  $i$ ,  $\vec{r}_i$  is the position vector of the atom with respect to the center of mass, and  $\vec{d}_i$  is the bond dipole vector.<sup>23</sup> As described by Bader *et al.*,<sup>23</sup> the contribution of the bond dipole vector  $\vec{d}_i$  to the molecular dipole is often small compared to  $q_i \vec{r}_i$ . From the MBIS calculation, we can obtain the atomic dipole moments, which can then be projected onto the bonds of each atom to calculate the bond dipole vector. In DASH, the bonds are distinguished by the attention order; thus, we can use the bond to the first atom in the DASH tree (highest attention) as the first projection vector and so forth.

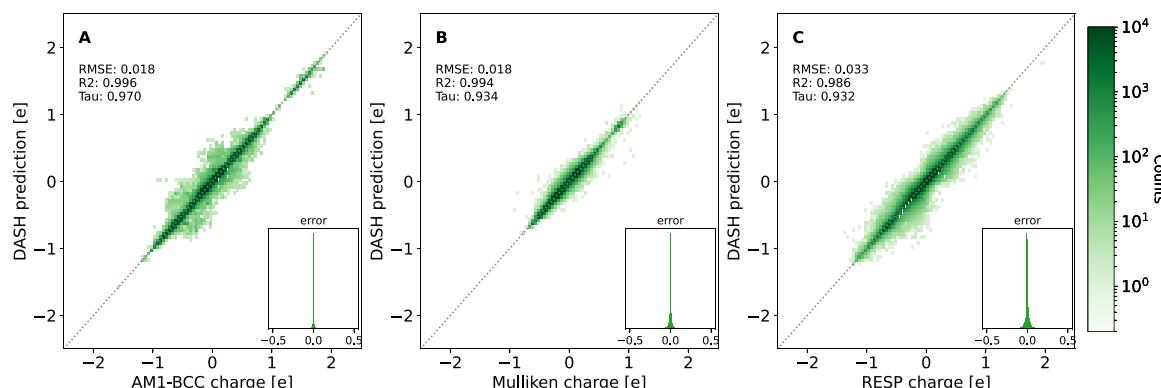
Again, the molecular dipole is experimentally measurable, but we compare the values estimated from the DASH tree with density functional theory (DFT) reference values due to the known deviations between QM and experiment.<sup>22</sup>

## E. Optimization of the DASH implementation

The implementation of the DASH tree published in Ref. 6 was adapted in this work to reduce runtime and to allow for an arbitrary number of additional properties. Pre-calculation of atom features and bond relations as well as a more efficient search for possible new subgraph expansions reduced the matching time by an order of magnitude compared to the original implementation. A separation between the tree structure and the stored data on a data structure and file structure allows for the addition of many properties. The improvements and code changes are described in more detail in the supplementary material.

## III. RESULTS AND DISCUSSION

In this work, we aim to test the hypothesis that the DASH tree built to assign MBIS partial charges captures sufficient general information about the environment of the atoms that the tree is directly applicable to the prediction of other atomic properties. The same data splits as in Lehner *et al.*<sup>6</sup> were used to compare the new properties with the corresponding reference values. Note that in contrast to Ref. 6, the validation set was not used for any meta-parameter optimization in this work. The results for the validation set are shown in



**FIG. 2.** Comparison between the computed reference values and the predictions with the DASH tree for partial charges of the validation set calculated with three different models: (a) AM1-BCC,<sup>11</sup> (b) Mulliken,<sup>10</sup> and (c) RESP.<sup>12</sup> The insets show a histogram of the error (difference between the reference and DASH values) without logarithmic scaling.

Secs. III A–III D, while those for the VEHICLE<sup>8</sup> test set are given in the supplementary material.

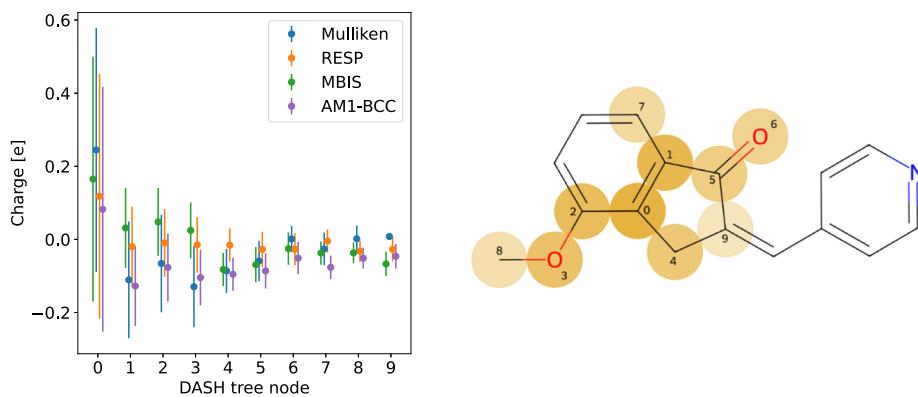
### A. Different partial charge models

The atomic properties most closely related to MBIS partial charges are other charge models such as AM1-BCC,<sup>11</sup> Mulliken,<sup>10</sup> or RESP.<sup>12</sup> Given the similarity, we expect that the DASH tree performs well for these, as can be seen in Fig. 2.

All three charge models show small root-mean-square-errors (RMSEs) and high correlations (as measured by  $R^2$  and Kendall's tau) between the partial charges assigned with the DASH tree and the reference values. 90% of predictions are found within 0.01e for AM1-BCC, 0.02e for Mulliken, and 0.04e for RESP. The very good performance of the DASH tree on the AM1-BCC and RESP charges—both used in classical force fields for molecular dynamics (MD) simulations—combined with the reproducibility, high speed,

and explainability of the approach renders DASH a flexible tool for generating partial charges for different simulation setups. Although the Mulliken charges are not as popular in MD, they are widely used as descriptors in machine learning (ML) approaches, where the DASH tree could be used to assign charges without the need for generating 3D conformers and performing QM calculations.

A comparison of all four different partial charge models for an example molecule is provided in Fig. 3. The different charges are calculated with the DASH tree for the atom highlighted in dark green in the right panel for an increasing matching depth in the tree. The left plot shows how all charge models converge to their final values during the matching process, while their prediction error decreases as more atoms are added to the subgraph. The added atoms are indicated with different shadings from dark yellow (initial) to light yellow (last added atom) in the right panel. These types of figures can be easily generated with the DASH tree and can serve as a source of explainability and error estimation.



**FIG. 3.** Example of the charge assignment process with the DASH tree for the dark yellow atom of the molecule shown on the right (atom 0). By expanding the subgraph from the initial atom, atoms in lighter yellow and higher number are added, and the charge predictions begin to converge to the final value, with decreasing assignment uncertainty (error bars).

## B. Atomic dispersion and polarizability

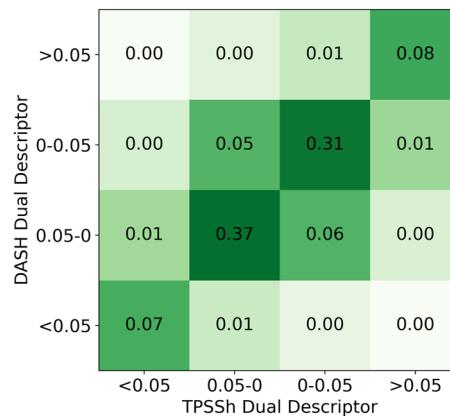
Next, the same DASH tree was used to predict the dispersion ( $C_6$ ) and polarizability of atoms. The reference values were calculated with DFT-D4.<sup>13</sup> As shown in Fig. 4, the DASH tree performs exceptionally well in predicting these properties.

## C. Electrophilicity and nucleophilicity

An important metric for the prediction of possible reactions of small molecules are Fukui functions.<sup>24–26</sup> They are a measure of the change in the electron density of a molecule when an electron is added or removed. Morell *et al.*<sup>15</sup> proposed to merge the Fukui functions into a single function termed the dual descriptor, which captures the same properties in a single, more robust function.<sup>27</sup> The dual descriptor is typically calculated as a function of the position in space relative to the molecule and visualized as a surface. Here, the dual descriptor function was integrated for each atom and the resulting parameter can be predicted with the DASH tree. The integration drastically reduces the resolution of the property since it is possible that the same atom has both strong electrophilic and nucleophilic sites, as seen in the simple case of the oxygen atom in water (see Fig. S4 of the [supplementary material](#)). Due to the loss in resolution, the task was modified into a classification problem to distinguish between weak and strong nucleophiles or electrophiles. We defined four bins as follows: strong nucleophiles ( $< -0.05$ ), weak nucleophiles ( $-0.05–0$ ), weak electrophiles ( $0–0.05$ ), and strong electrophiles ( $>0.05$ ). As can be seen in Fig. 5, there is a good performance by the DASH tree overall, with some confusion between weak nucleophiles and weak electrophiles.

## D. Molecular properties

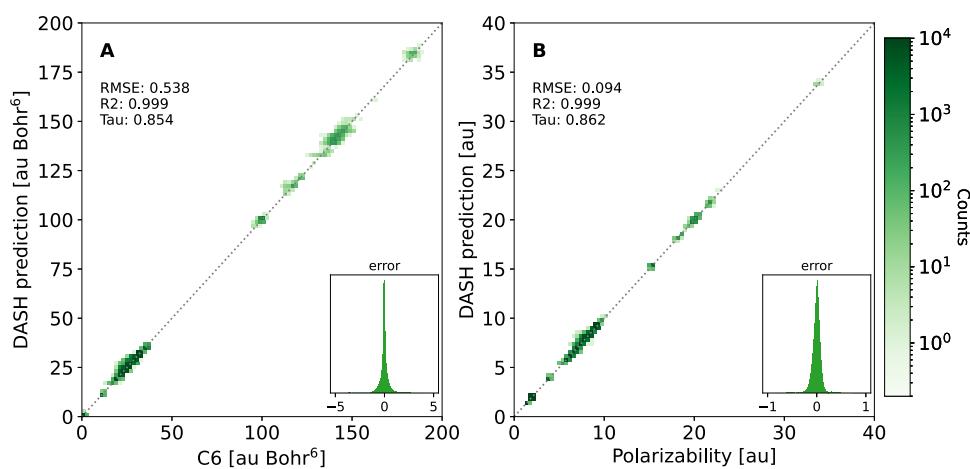
The polarizability of a molecule was calculated with Eq. (2) and compared to the reference values [Fig. 6(a)]. The DASH approach shows a high correlation and small RMSE compared to the



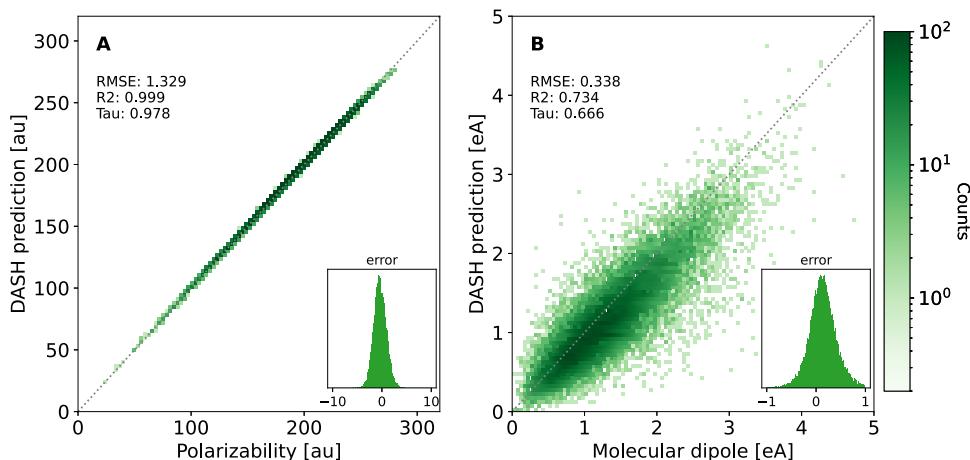
**FIG. 5.** Confusion matrix of the integrated dual descriptor calculated with Psi4<sup>17</sup> and predicted by the DASH tree for the validation set. Four bins were considered: strong nucleophiles ( $< -0.05$ ), weak nucleophiles ( $-0.05–0$ ), weak electrophiles ( $0–0.05$ ), and strong electrophiles ( $>0.05$ ).

reference values. Kendall's tau is even better than for the atomic polarizabilities, possibly due to some cancellation of errors.

The molecular dipole moment was calculated from the atomic contributions with the function described by Bader *et al.*<sup>23</sup> [Eq. (3)]. Although the strength of the atomic dipole predicted with DASH has a low correlation to its reference (Fig. S2 of [supplementary material](#)), the molecular dipole predicted using DASH shows a good correlation to the reference values [Fig. 6(b)]. The relatively poor performance of DASH for the atomic contributions comes from the fact that DASH is based on the 2D topology of the molecules and the strong 3D conformational dependency of the atomic dipole strength is ignored. Thus, for high prediction accuracy, a model would need to include 3D information, as shown, for example, in Ref. 28 with a graph neural network.



**FIG. 4.** Comparison of the atomic dispersion (a) and polarizability (b) between DFT-D4<sup>13</sup> and the DASH tree for the validation set. The insets show a histogram of the error (difference between the reference and predicted values) without logarithmic scaling.



**FIG. 6.** Comparison of molecular polarizability (a) and the molecular dipole moment (b) between reference values and the predictions with the DASH tree for the validation set. The insets show a histogram of the error (difference between the reference and predicted values) without logarithmic scaling.

#### IV. CONCLUSIONS

We report that the DASH tree developed to predict MBIS partial charges can be exploited to predict other atomic properties, without re-training the graph neural network from which the tree was built. Thus, the substructures in the DASH tree are shown to be transferable between properties and encode the important local electronic environment of an atom. We demonstrated this by predicting different partial charge models (AM1-BCC, Mulliken, and RESP), atomic dispersion ( $C_6$ ), atomic polarizability, and dual descriptor for electrophilicity and nucleophilicity. Furthermore, molecular properties, such as molecular polarizability and molecular dipole, are accessible if there is a relationship with atomic contributions predicted by the DASH tree. Even if the atomic property is negatively impacted by the 2D nature of the description in DASH (i.e., the 3D conformational dependence is ignored), as seen for atomic dipole moment, a good performance was found for the resulting molecular dipole.

Predicting AM1-BCC and RESP charges with the DASH tree can be beneficial for existing fixed-charge force fields, as it provides a computationally efficient and user-friendly method. As the DASH approach works with 2D structures, conformational averaging is not required and the runtime scales linearly with the number of atoms in a molecule instead of the higher-order scaling of semi-empirical methods or DFT methods. In addition, it provides high accuracy while maintaining a high level of explainability due to fragment-based assignment.

We conclude that the DASH tree can be used as an accurate, explainable, and computationally efficient model to predict a diverse set of atomic and molecular properties. It presents an attractive alternative for semi-empirical methods or ML approaches for property calculation, especially for applications with a large number of molecules. The final DASH tree with the different atomic properties as well as the complete dataset with wave functions is made freely available.

#### SUPPLEMENTARY MATERIAL

See the [supplementary material](#) for details on implementation optimizations, additional information on the atomic dipole moment and dual descriptor, and the results for the VEHICLE test set.

#### ACKNOWLEDGMENTS

The numerical simulations were performed on the high performance computing cluster Euler by ETH Zurich. The authors gratefully acknowledge financial support by ETH Zurich (Grant No. ETH-50 21-1).

#### AUTHOR DECLARATIONS

##### Conflict of Interest

The authors have no conflicts to disclose.

##### Author Contributions

**Marc T. Lehner:** Conceptualization (lead); Data curation (lead); Investigation (lead); Methodology (lead); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead). **Paul Katzberger:** Conceptualization (supporting); Data curation (supporting); Investigation (supporting); Methodology (supporting); Software (supporting); Validation (supporting); Visualization (supporting); Writing – original draft (supporting). **Niels Maeder:** Conceptualization (supporting); Data curation (supporting); Investigation (supporting); Methodology (supporting); Software (supporting); Validation (supporting); Visualization (supporting); Writing – original draft (supporting). **Gregory A. Landrum:** Conceptualization (supporting); Data curation (supporting); Investigation (supporting); Methodology (supporting);

Software (supporting); Validation (supporting); Writing – review & editing (supporting). **Sereina Riniker:** Conceptualization (equal); Funding acquisition (equal); Methodology (supporting); Project administration (lead); Resources (lead); Supervision (lead); Writing – review & editing (equal).

## DATA AVAILABILITY

The implementation of the DASH tree and all software used for this work are open source and available on GitHub (<https://github.com/rinikerlab/DASH-tree>). The dataset with the atomic and molecular properties is freely available via the ETH Research Collection (<https://www.research-collection.ethz.ch/handle/20.500.11850/670546>). The logistics of providing the full dataset of all 3D structures and wave functions (55 TB) are more complex. We are making these data freely available via Globus ([app.globus.org](http://app.globus.org)). Please contact the corresponding author for information on how to access it.

## REFERENCES

- <sup>1</sup>B. K. Rai and G. A. Bakken, “Fast and accurate generation of *ab initio* quality atomic charges using nonparametric statistical regression,” *J. Comput. Chem.* **34**, 1661–1671 (2013).
- <sup>2</sup>P. Bleiziffer, K. Schaller, and S. Riniker, “Machine learning of partial charges derived from high-quality quantum-mechanical calculations,” *J. Chem. Inf. Model.* **58**, 579–590 (2018).
- <sup>3</sup>R. Martin and D. Heider, “ContraDRG: Automatic partial charge prediction by machine learning,” *Front. Genet.* **10**, 990 (2019).
- <sup>4</sup>K. Kato, T. Masuda, C. Watanabe, N. Miyagawa, H. Mizouchi, S. Nagase, K. Kamisaka, K. Oshima, S. Ono, H. Ueda, A. Tokuhisa, R. Kanada, M. Ohta, M. Ikeguchi, Y. Okuno, K. Fukuzawa, and T. Honma, “High-precision atomic charge prediction for protein systems using fragment molecular orbital calculation and machine learning,” *J. Chem. Inf. Model.* **60**, 3361–3368 (2020).
- <sup>5</sup>Y. Wang, I. Pulido, K. Takaba, B. Kaminow, J. Scheen, L. Wang, and J. D. Chodera, “EspalomaCharge: Machine learning-enabled ultrafast partial charge assignment,” *J. Phys. Chem. A* **128**, 4160–4167 (2024).
- <sup>6</sup>M. T. Lehner, P. Katzberger, N. Maeder, C. C. G. Schiebroek, J. Teetz, G. A. Landrum, and S. Riniker, “DASH: Dynamic attention-based substructure hierarchy for partial charge assignment,” *J. Chem. Inf. Model.* **63**, 6014–6028 (2023).
- <sup>7</sup>T. Verstraelen, S. Vandenbrande, F. Heidar-Zadeh, L. Vanduyfhuys, V. Van Speybroeck, M. Waroquier, and P. W. Ayers, “Minimal basis iterative stockholder: Atoms in molecules for force-field development,” *J. Chem. Theory Comput.* **12**, 3894–3912 (2016).
- <sup>8</sup>W. R. Pitt, D. M. Parry, B. G. Perry, and C. R. Groom, “Heteroaromatic rings of the future,” *J. Med. Chem.* **52**, 2952–2963 (2009).
- <sup>9</sup>S. Boothroyd, P. K. Behara, O. C. Madin, D. F. Hahn, H. Jang, V. Gapsys, J. R. Wagner, J. T. Horton, D. L. Dotson, M. W. Thompson, J. Maat, T. Gokey, L.-P. Wang, D. J. Cole, M. K. Gilson, J. D. Chodera, C. I. Bayly, M. R. Shirts, and D. L. Mobley, “Development and benchmarking of open force field 2.0.0: The Sage small molecule force field,” *J. Chem. Theory Comput.* **19**, 3251–3275 (2023).
- <sup>10</sup>R. S. Mulliken, “Electronic population analysis on LCAO-MO molecular wave functions. I,” *J. Chem. Phys.* **23**, 1833–1840 (1955).
- <sup>11</sup>A. Jakalian, D. B. Jack, and C. I. Bayly, “Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation,” *J. Comput. Chem.* **23**, 1623–1641 (2002).
- <sup>12</sup>C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model,” *J. Phys. Chem.* **97**, 10269–10280 (1993).
- <sup>13</sup>E. Caldeweyher, C. Bannwarth, and S. Grimme, “Extension of the D3 dispersion coefficient model,” *J. Chem. Phys.* **147**, 034112 (2017).
- <sup>14</sup>K. Fukui, “Role of Frontier orbitals in chemical reactions,” *Science* **218**, 747–754 (1982).
- <sup>15</sup>C. Morell, A. Grand, and A. Toro-Labbé, “New dual descriptor for chemical reactivity,” *J. Phys. Chem. A* **109**, 205–212 (2005).
- <sup>16</sup>M. T. Lehner, P. Katzberger, and S. Riniker (2023). “QM dataset for MBIS partial charges,” ETH Research Collection. <https://doi.org/10.3929/ethz-b-000613415>
- <sup>17</sup>J. M. Turney, A. C. Simonett, R. M. Parrish, E. G. Hohenstein, F. A. Evangelista, J. T. Fermann, B. J. Mintz, L. A. Burns, J. J. Wilke, M. L. Abrams, N. J. Russ, M. L. Leininger, C. L. Janssen, E. T. Seidl, W. D. Allen, H. F. Schaefer, R. A. King, E. F. Valeev, C. D. Sherrill, and T. D. Crawford, “Psi4: An open-source *ab initio* electronic structure program,” *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2**, 556–565 (2012).
- <sup>18</sup>J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general Amber force field,” *J. Comput. Chem.* **25**, 1157–1174 (2004).
- <sup>19</sup>D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, M. K. Gilson, and P. K. Eastman, “Escaping atom types in force fields using direct chemical perception,” *J. Chem. Theory Comput.* **14**, 6076–6092 (2018).
- <sup>20</sup>D. A. Case, H. M. Aktulga, K. Belfon, I. Y. Ben-Shalom, J. T. Berryman, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, G. A. Cisneros, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, G. Giambasu, M. K. Gilson, H. Gohlke, A. W. Goetz, R. Harris, S. Izadi, S. A. Izmailov, K. Kasavajhala, M. C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K. M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K. A. O’Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C. L. Simmerling, N. R. Skrynnikov, J. Smith, J. Swails, R. C. Walker, J. Wang, J. Wang, H. Wei, R. M. Wolf, X. Wu, Y. Xiong, Y. Xue, D. M. York, S. Zhao, and P. A. Kollman, AMBER 2022, University of California, San Francisco, 2022.
- <sup>21</sup>C. Bannwarth, S. Ehlert, and S. Grimme, “GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions,” *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
- <sup>22</sup>A. L. Hickey and C. N. Rowley, “Benchmarking quantum chemical methods for the calculation of molecular dipole moments and polarizabilities,” *J. Phys. Chem. A* **118**, 3678–3687 (2014).
- <sup>23</sup>R. Bader, A. Larouche, C. Gatti, M. Carroll, P. MacDougall, and K. Wiberg, “Properties of atoms in molecules: Dipole moments and transferability of properties,” *J. Chem. Phys.* **87**, 1142–1152 (1987).
- <sup>24</sup>R. G. Parr and W. Yang, “Density functional approach to the frontier-electron theory of chemical reactivity,” *J. Am. Chem. Soc.* **106**, 4049–4050 (1984).
- <sup>25</sup>W. Langenaeker, M. De Decker, P. Geerlings, and P. Raeymaekers, “Quantum-chemical study of the Fukui function as a reactivity index: Probing the acidity of bridging hydroxyls in zeolite-type model systems,” *J. Mol. Struct.: THEOCHEM* **207**, 115–130 (1990).
- <sup>26</sup>P. W. Ayers, W. T. Yang, and L. J. Bartolotti, “Fukui function,” in *Chemical Reactivity Theory: A Density Functional View*, edited by P. K. Chattaraj (CRC Press, 2009).
- <sup>27</sup>J. I. Martínez-Araya, “Why is the dual descriptor a more accurate local reactivity descriptor than fukui functions?,” *J. Math. Chem.* **53**, 451–465 (2015).
- <sup>28</sup>M. Thürlemann, L. Bösel, and S. Riniker, “Learning atomic multipoles: Prediction of the electrostatic potential with equivariant graph neural networks,” *J. Chem. Theory Comput.* **18**, 1701–1710 (2022).
- <sup>29</sup>D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” *J. Chem. Inf. Model.* **50**, 742–754 (2010).