This article appeared in J. Chem. Phys. **155**, 204111 (2021) and may be found at https://aip.scitation.org/doi/10.1063/5.0055393.

Impact of quantum-chemical metrics on the machine learning prediction of electron density

Ksenia R. Briling,[1] Alberto Fabrizio,[1,2] and Clemence Corminboeuf[1,2,a)]
[1)]*Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*
[2)]*National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

(Dated: 13 July 2021)

Machine learning (ML) algorithms have undergone an explosive development impacting every aspect of computational chemistry. To obtain reliable predictions, one needs to maintain the proper balance between the black-box nature of ML frameworks and the physics of the target properties. One of the most appealing quantum-chemical properties for regression models is the electron density, and some of us recently proposed a transferable and scalable model based on the decomposition of the density onto an atom-centered basis set. The decomposition, as well as the training of the model, is at its core a minimization of some loss function, which can be arbitrarily chosen and may lead to results of different quality. Well-studied in the context of density fitting (DF), the impact of the metric on the performance of ML models has not been analyzed yet. In this work, we compare predictions obtained using the overlap and the Coulomb-repulsion metrics for both decomposition and training. As expected, the Coulomb metric used as both the DF and ML loss functions leads to the best results for the electrostatic potential and dipole moments. The origin of this difference lies in the fact that the model is not constrained to predict densities that integrate to the exact number of electrons N. Since an *a posteriori* correction for the number of electrons decreases the errors, we proposed a modification of the model where N is included directly into the kernel function, which allowed to lower the errors on the test and out-of-sample sets.

I. INTRODUCTION

The molecular electron density $\rho(\mathbf{r})$ is one of the cornerstones of modern quantum chemistry and chemical physics. Unlike the many-body wavefunction, the electron density, being a much simpler real-space scalar function, is an observable and can be measured by X-ray diffraction[1] or transmission electron microscopy.[2] At the same time, as shown by the first Hohenberg–Kohn theorem,[3] $\rho(\mathbf{r})$ embodies the same information as the wavefunction, and thus gives access to all molecular properties either directly or from its deformations in the presence of external fields. Because of its fundamental role in electronic structure theory, the electron density is a highly appealing target for machine learning (ML) models, which is demonstrated by the growing number of works on the non-linear regression of $\rho(\mathbf{r})$. These models can be divided in two categories: those treating the field as a set of values on a real-space grid[4–7] and those built on a decomposition onto a basis set.[8]

In this second category, we have recently developed[9,10] and demonstrated the wide-scope applicability[10,11] and the generality[12] of a transferable model of the electron density. The model is based on symmetry-adapted Gaussian process regression (SA-GPR)[13,14] and on a local decomposition of the electron density field into an atom-centered spherical Gaussian basis. The decomposition, as any density-fitting (DF) approximation, consists of two critical parts: the selection[10] or construction[12] of a suitable basis set and the determination of the basis set expansion coefficients. The coefficients are determined by minimizing a loss function between the fitted and the *ab initio* densities. The set of density-decomposition coefficients represent the target of the machine learning model. During the training phase, the regression weights are found by minimizing a second loss function, which reflects the difference between the decomposed density and the predicted one.

In principle, any function of a set of real-space variables, such as the electron density, can be exactly expanded onto a complete set of basis functions in a unique way. In practice, the auxiliary basis sets are incomplete and the use of different loss functions leads to different expansions of the electron density.

The simplest way to fit the approximate density to the original $\rho(\mathbf{r})$ is to apply the least-squares technique[15–19] and find the decomposition coefficients $\{c_i^{\mathrm{DF}}\}$ that minimize the error

$$\text{fitting error} = \int \left|\rho(\mathbf{r}) - \sum_i c_i^{\mathrm{DF}} \phi_i(\mathbf{r})\right|^2 \mathrm{d}^3\mathbf{r}. \qquad (1)$$

This intuitive form of the fitting (decomposition) loss function can be re-stated (usually under the constraint for the number of electrons) in a more general quadratic functional of the density residue $\Delta\rho(\mathbf{r}) = \rho(\mathbf{r}) - \sum_i c_i^{\mathrm{DF}}\phi_i(\mathbf{r})$,

$$\text{fitting error} = \iint \Delta\rho(\mathbf{r}_1)\,\hat{\mathrm{O}}(\mathbf{r}_1,\mathbf{r}_2)\Delta\rho(\mathbf{r}_2)\,\mathrm{d}^3\mathbf{r}_1\,\mathrm{d}^3\mathbf{r}_2, \qquad (2)$$

where $\hat{\mathrm{O}}(\mathbf{r}_1,\mathbf{r}_2)$ is a two-electron operator. Eq. 1 is a special case of Eq. 2 where $\hat{\mathrm{O}}$ is the overlap operator $\hat{\mathrm{S}} = \delta(\mathbf{r}_{12})$.

The overlap fitting yields approximate densities that often lack accuracy.[19–22] For this reason, nowadays the standard procedure for density-fitting applications is the electrostatic repulsion fitting[23–28] with $\hat{\mathrm{O}}$ being the Coulomb operator $\hat{\mathrm{J}} = |\mathbf{r}_{12}|^{-1}$, which gives an approximate density whose electric field is the closest to the original one.

a)Electronic mail: clemence.corminboeuf@epfl.ch

This article appeared in J. Chem. Phys. **155**, 204111 (2021) and may be found at https://aip.scitation.org/doi/10.1063/5.0055393.

The generality of Eq. 2 promotes other ways to find the decomposition coefficients. For example, the anti-Coulomb metric[29] $\hat{\mathrm{O}} = -|\mathbf{r}_{12}|$, although not widely used, gives an approximate density with the closest electrostatic potential to the reference. For extended systems, in order to avoid the slow decay of the Coulomb operator, the complementary error-function Coulomb metric[22,30] $\hat{\mathrm{O}} = \mathrm{erfc}(\omega|\mathbf{r}_{12}|)\cdot|\mathbf{r}_{12}|^{-1}$ and the Gaussian-damped Coulomb metric[30] $\hat{\mathrm{O}} = \exp(-\omega|\mathbf{r}_{12}|^2)\cdot|\mathbf{r}_{12}|^{-1}$ were also proposed. Since both of them provide a smooth transition from the Coulomb ($\omega \to 0$) to the scaled overlap ($\omega \to \infty$) metric, for our purposes it is sufficient to consider only the two limiting cases.

For the same reasons as the fitting of the *ab initio* electron density, the choice of the loss function to fix the regression weights is also not unique. In fact, a simple regression model of the electron density can also be formulated as a least-squares problem, where the task is to find the regression weights $\{x_j\}$ that minimize a quadratic loss function

$$\Lambda(\mathbf{x}) = \sum_{\substack{\text{training}\\ \text{set}}} \int \left|\rho(\mathbf{r}) - \sum_i c_i^{\mathrm{ML}}(\mathbf{x})\phi_i(\mathbf{r})\right|^2 \mathrm{d}^3\mathbf{r}. \qquad (3)$$

Eq. 3 has the same structure as the overlap density-fitting problem of Eq. 1 and can be generalized in the same fashion as Eq. 2. The possibility to change the predicted expansion coefficients simply by changing the metric both in the initial density decomposition and in the regression loss function allows, in principle, the tuning of the SA-GPR machinery for each specific application of the predicted electron density.

In principle, it is possible to construct ML loss functions using also other integrals targeted to DFT energies and energy densities, *e.g.* containing $\Delta\rho^{4/3}$ or reduced density gradients. However, such loss functions cannot be written as quadratic functions of the regression weights, and the learning step would require an iterative solution in a self-consistent manner.

Our previous works targeting the electron density with SA-GPR coincidentally exploited two different metrics ($\hat{\mathrm{S}}$[9,12] and $\hat{\mathrm{J}}$[10]) for the decomposition but only the overlap metric in the machine learning loss function. Given the known effects of the metric choice in the density-fitting literature[19–22] and the lack of a corresponding systematic analysis for machine learning applications, many questions remain unanswered. For instance, is the density decomposed with one metric more difficult to learn than another? Do the associated predicted densities differ significantly? Which combinations of loss functions are the most efficient for which application? More generally, these questions also address a perhaps more fundamental topic that is how do ML models interact with deductive reasoning?

In the present work, we apply the four possible combinations of $\hat{\mathrm{S}}$ and $\hat{\mathrm{J}}$ metrics on the same set of biologically-relevant molecules and compare the quality of the predicted electron density to reproduce different electronic properties, ranging from the number of electrons to the dipole moments, electrostatic potentials (ESP), and the characterization of the intra- and intermolecular electronic fingerprints with the density overlap region indicator (DORI).[31] As a result of this systematic analysis, we also introduce several different schemes to restore the correct number of electrons in the predicted electron densities.

II. COMPUTATIONAL DETAILS

This work uses the side-chain–side-chain interaction subset of the BioFragment database[32] (BFDb). From the original set, we excluded molecules containing sulfur atoms and/or more than 25 atoms, as well as several structures with unphysical atomic distances. The final dataset contains 2287 dimers and 35 of the most representative monomer structures. Out of the total set, 2000 structures (1975 dimers and 25 monomers) were randomly selected for the training set and 322 structures for the test set.

All quantum-chemical computations, except for three- and four-center overlap integrals, were made with a locally modified version of PySCF.[33,34] The reference density matrices were computed at the ωB97X-D[35]/cc-pVQZ[36] level with the RI-JK approximation. For density decomposition, the cc-pVQZ/JKFIT[37] basis was used.

For sampling electrostatic potentials, we computed molecular surfaces[38] $p(\mathbf{r}) = p_0$ with

$$p(\mathbf{r}) = \iint s(\mathbf{r}-\mathbf{r}_1)\rho_1(\mathbf{r}_1,\mathbf{r}_2)s(\mathbf{r}-\mathbf{r}_2)\,\mathrm{d}^3\mathbf{r}_1\,\mathrm{d}^3\mathbf{r}_2, \qquad (4)$$

where $\rho_1(\mathbf{r}_1,\mathbf{r}_2)$ is the *ab initio* one-particle density matrix, $s(\mathbf{r}) = \exp(-a|\mathbf{r}|^2)$, $a = 1/16$, and $p_0 = 1/16, 1/4, 1, 4$, or 32. The error in the predicted electrostatic potential $U_{\mathrm{ML}}(\mathbf{r})$ with respect to the *ab initio* one $U(\mathbf{r})$ is defined as

$$\varepsilon_{\mathrm{ESP}} = \sqrt{\frac{\iint_S (U(\mathbf{r}) - U_{\mathrm{ML}}(\mathbf{r}))^2\,\mathrm{d}S}{\iint_S \mathrm{d}S}}, \qquad (5)$$

and the surfaces are discretized with the spherical quadrature rules.[39,40]

The density overlap region indicator[31] was computed analytically on a cubic grid with a spacing of 0.1 Bohr. The comparison between two DORI fields in real space was done using the Walker–Mezey similarity measure[41] $L(a,a')$ with $(a,a') = (0.1,0.7)$, $(0.7,0.95)$, and $(0.95,1)$.

The error in the predicted dipole moment $\boldsymbol{\mu}_{\mathrm{ML}}$ with respect to the *ab initio* one $\boldsymbol{\mu}$ is defined as

$$\varepsilon_{\mathrm{dipole}} = |\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathrm{ML}}|. \qquad (6)$$

For a density $\rho'(\mathbf{r})$, we define the absolute

$$E_O[\rho'|\rho] = (\rho' - \rho|\,\hat{\mathrm{O}}\,|\rho' - \rho) \qquad (7)$$

and relative

$$e_O[\rho'|\rho] = E_O[\rho'|\rho]/(\rho|\,\hat{\mathrm{O}}\,|\rho) \qquad (8)$$

errors with respect to $\rho(\mathbf{r})$ to be consistent with density-fitting and machine learning loss functions.

The tensorial λ-SOAP kernels[13,14] were computed with the following parameters: environment cutoff $r_{\mathrm{cut}} = 4$ Å, Gaussian smearing $\sigma = 0.3$ Å, angular cutoff $l_{\mathrm{cut}} = 6$, radial cutoff

$n_{\text{cut}} = 8$, environmental kernel exponent $\zeta = 2$. A subset of $M = 1000$ reference environments was taken to reduce the dimensionality of the regression problem, and the regularization parameter η was set to 10^{-6}.

III. THE QUANTUM-CHEMICAL METRICS

A. Model construction

Building a ML model for the electron density first consists in fitting a linear combination of atom-centered basis functions $\{\phi_i\}$

$$\rho_{\text{DF}}(\mathbf{r}) = \sum_i c_i^{\text{DF}} \phi_i(\mathbf{r}) \tag{9}$$

to the molecular electron density $\rho_{\text{QM}}(\mathbf{r})$, which can be written in terms of the one-electron density matrix or computed on a real-space grid etc. The fitting coefficients $\{c_i^{\text{DF}}\}$ are chosen to minimize a density-fitting (DF) loss function

$$\Lambda_{\text{DF}}(\mathbf{c}^{\text{DF}}) = (\rho_{\text{DF}} - \rho_{\text{QM}} | \hat{\text{O}} | \rho_{\text{DF}} - \rho_{\text{QM}}) \to \min, \tag{10}$$

where $\hat{\text{O}}$ is a two-electron operator (overlap $\hat{\text{S}} = \delta(|\mathbf{r}_{12}|)$ or Coulomb repulsion $\hat{\text{J}} = |\mathbf{r}_{12}|^{-1}$) and the shorthand for two-electron integrals is

$$(f | \hat{\text{O}} | g) = \iint f(\mathbf{r}_1)\, \hat{\text{O}}(\mathbf{r}_1, \mathbf{r}_2) g(\mathbf{r}_2)\, \text{d}^3\mathbf{r}_1\, \text{d}^3\mathbf{r}_2. \tag{11}$$

The solution for Eq. 10 is

$$\mathbf{c}^{\text{DF}} = \mathbf{O}^{-1}\mathbf{w}, \tag{12}$$

where $O_{ij} = (\phi_i | \hat{\text{O}} | \phi_j)$ are the matrix elements of the operator $\hat{\text{O}}$ and $w_i = (\phi_i | \hat{\text{O}} | \rho_{\text{QM}})$ are, in the case of $O = S$, the projections of the target field ρ_{QM} onto the decomposition basis $\{\phi_i\}$. Different operators $\hat{\text{O}}$ yield different sets of coefficients $\{c_i^{\text{DF}}\}$, each of which minimizes the loss function associated with $\hat{\text{O}}$.

In the same spirit, the ML loss function can be also written as a sum over the structures of the training set (TrS)

$$\Lambda_{\text{ML}}(\mathbf{x}) = \sum_{\text{TrS}} (\rho_{\text{ML}} - \rho_{\text{DF}} | \hat{\text{O}}' | \rho_{\text{ML}} - \rho_{\text{DF}}) \to \min, \tag{13}$$

where each "predicted" density

$$\rho_{\text{ML}}(\mathbf{r}) = \sum_i c_i^{\text{ML}}(\mathbf{x}) \phi_i(\mathbf{r}) \tag{14}$$

depends on the regression weights $\mathbf{x}$ via a kernel function

$$\mathbf{c}^{\text{ML}}(\mathbf{x}) = \mathbf{K}\mathbf{x}, \tag{15}$$

and $\hat{\text{O}}'$ is also a two-particle operator.

The DF metric O and the ML metric O' are independent and, in principle, can be chosen to be different. For example, as we did in Ref. 10, it is perfectly possible to take $\hat{\text{O}} = \hat{\text{J}}$ so that the decomposed densities ρ_{DF} are the closest to the *ab initio* densities ρ_{QM} in the sense that the self-repulsion of their residuals is the minimum, and then take $\hat{\text{O}}' = \hat{\text{S}}$ so that the training-set predictions are (on average) the closest to ρ_{DF} in the sense that their overlap is the maximum.

However, the use of different metrics for O and O' has a formally unclear physical meaning. On the other hand, using the same metric O at both DF and ML steps is analogous to the minimization of a loss function

$$\Lambda_{\text{DF+ML}}(\mathbf{x}) = \sum_{\text{TrS}} (\rho_{\text{ML}} - \rho_{\text{QM}} | \hat{\text{O}} | \rho_{\text{ML}} - \rho_{\text{QM}}), \tag{16}$$

making the predictions to be the closest to the original density in the O-sense, as we did with the S-metric in Ref. 9.

B. Results

In this work, four sets of densities were predicted from the four possible combinations of DF and ML metrics OO': JJ, JS, SJ, and SS. Comparison of the electron density mean errors $E_O[\rho|\rho_{\text{QM}}]$ (*i.e.*, with respect to ρ_{QM}) for the two sets of fitted densities and four sets of predictions can be found in Fig. 1a and 1b. Among the predictions, the lowest J-error is observed (on average) for the JJ-scheme, where the J-metric is used for both DF and ML steps; the SS-scheme, where the J-metric is not used at all, gives the highest J-error. When using S-errors, the ranking is opposite. It is not surprising, because the goal of the framework is to yield the optimal predicted densities, and what is optimal is defined by the DF and ML metrics.

However, while the differences within the S-errors are less than 6%, the J-errors difference goes up to an order of magnitude. It seems that the ML loss function with the S-metric has a more shallow minimum, which can be already expected from the analysis of errors in the fitted densities alone (see also Table S4 in the Supplementary Material): it is clear that the J-metric not only yields a smaller error in the number of electrons than the S-metric, but is also more sensitive to small density differences. It is interesting that the JS-scheme performs worse than the SJ, regardless of the error metric. More detailed analysis shows that the S- and J-errors are more sensitive to the DF and ML metric respectively. This results in the error for JS being larger than for SJ in both cases.

The learning curves (Fig. S1a of the Supplementary Material), which are based on the relative prediction errors (using the corresponding ML metrics) with respect to their reference fitted densities, show that the predictions are almost independent on the fitting metric (the curves for the JO- and SO-schemes are nearly the same, for both $O = S$ or J). This trend essentially means that both the S- and J-fitting coefficients, $\{c_i^{\text{DF}}\}$, correlate with the atomic representation in a similar way. It is known[22] that the contribution $\sum_j (\mathbf{O}^{-1})_{ij} (\phi_j | \hat{\text{O}} | \chi_p \chi_q)$ of an auxiliary function ϕ_i, centered on one atom, to the product of two basis functions χ_p and χ_q, centered on another atom, decays much slower with increasing distance between these two atoms for repulsion metric than for the overlap one due to long-rangeness of the $\hat{\text{J}}$ operator. Potentially it can make J-coefficients harder to learn,

This article appeared in J. Chem. Phys. **155**, 204111 (2021) and may be found at https://aip.scitation.org/doi/10.1063/5.0055393.

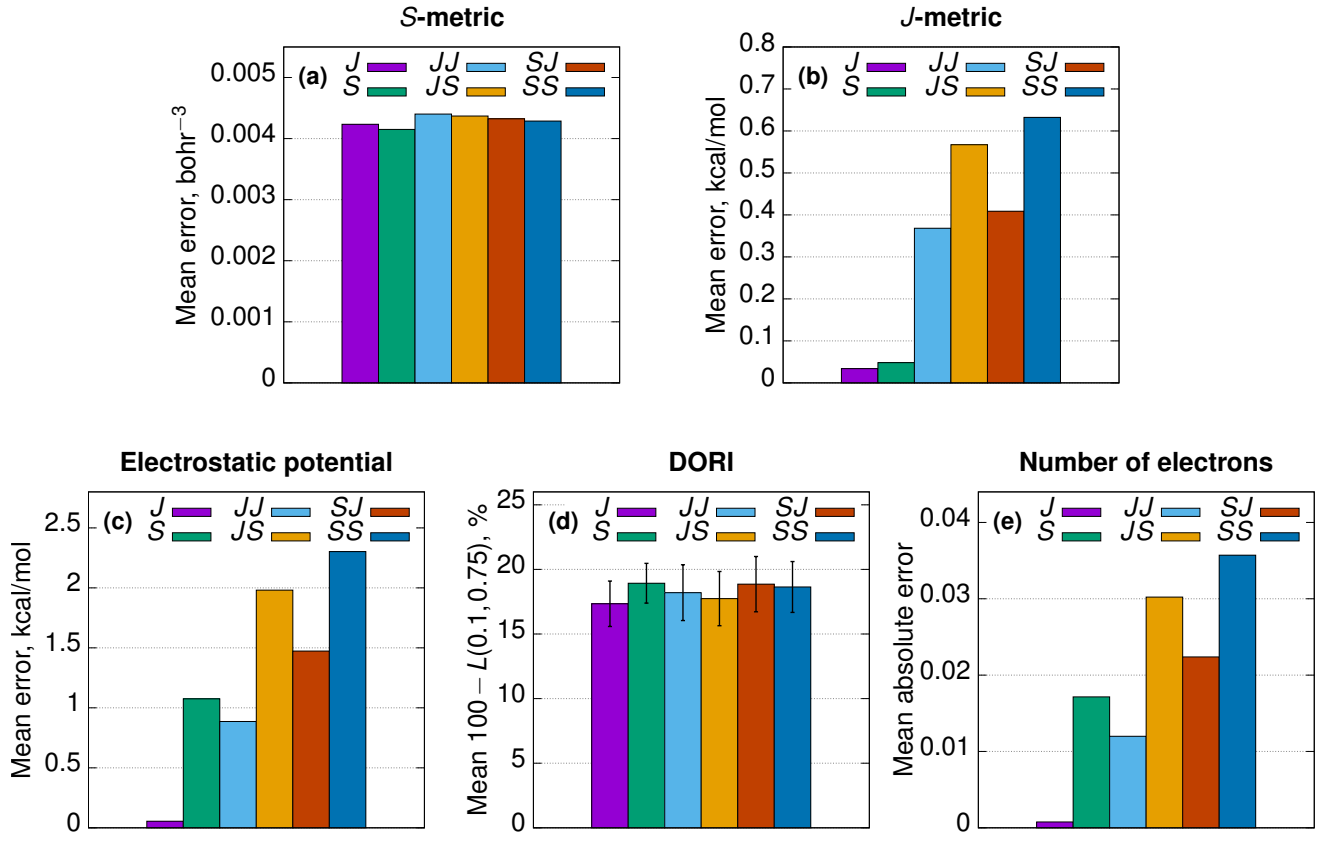

FIG. 1. Mean error measures computed on the test set for fitted (J and S) and predicted (JJ, JS, SJ, and SS) densities without any constraints on the number of electrons: *(a)* S- and *(b)* J- metrics computed as $(\rho_{\mathrm{ML}}-\rho_{\mathrm{QM}}|\hat{\mathrm{O}}|\rho_{\mathrm{ML}}-\rho_{\mathrm{QM}})$; *(c)* errors in the electrostatic potential on the isosurface $p_0 = 4$, corresponding to average density $\langle\langle\rho\rangle\rangle \approx 2\times10^{-6}\,\mathrm{bohr}^{-3}$; *(d)* Walker–Mezey similarity indices for DORI in the non-covalent region (between 0.1 and 0.75), error bars show the standard deviation; *(e)* absolute errors in the number of electrons. All errors are computed with respect to corresponding ρ_{QM}. Note that for the S-metric, we here use the L^2 and not the L^1 norm as in Ref. 10 (see Fig. S6 for comparison).

but in our set all the molecules were not big enough to make the difference in decay noticeable.

On the other hand, the two sets of superimposed curves (OS and OJ on Fig. S1a of the Supplementary Material) do differ. The fact that for the OS-curves the improvement from 250 to 2000 training molecules is slightly less significant than for the OJ-curves can be explained by the lower sensitivity of the S-metric as discussed above. Even though the OS-curves are lower than the OJ ones, it does not mean that the former predictions are in any way better: these are pure prediction errors with respect to the corresponding fitted densities, moreover, S-errors are shown for the OS-schemes and J-errors — for the OJ-schemes. In any case, the full-training-set prediction errors (the last points of the learning curves) are very close (5×10^{-5}% for OS and 6×10^{-5}% for OJ) and the difference is not important.

To test these metric combinations on real-life applications, we first chose two fundamentally different properties: the electrostatic potential and the density overlap region indicator computed on all the test set molecules. As shown in Fig. 1c and 1d, all four schemes lead to electrostatic potentials with a very different quality with the J-metric consistently decreasing the error. This result is not surprising considering that the J-metric yields an electron density whose electric field is the closest to the reference one.[24,25] Because of the slow decay of the Coulomb potential, the J-metric incorporates accurately long-range information into the expansion coefficients and it is generally preferred in common quantum-chemical applications.[19–22]

Since the model is not constrained to produce densities that integrate to the correct number of electrons N (see Sec. IV A), the errors $|\Delta N|$ are also shown on Fig. 1e. The quality of the electrostatic potential does correlate with the error in the number of electrons due to non-locality of both the properties. On the other hand, DORI is much less sensitive to $|\Delta N|$, because it explicitly depends only on the local wave vector $\nabla\rho(\mathbf{r})/\rho(\mathbf{r})$. As a simple example, a uniform scaling of a reference density by a factor of x leads to an error of $(x-1)N$ in the number of electrons and thus to an error in the electrostatic potential, but with no influence on DORI. Hence, the advantage of the J-metric observed above arises from the fact that it usually yields a smaller error in the number of electrons.

This article appeared in J. Chem. Phys. **155**, 204111 (2021) and may be found at https://aip.scitation.org/doi/10.1063/5.0055393.

Even for the test set, which is made of structures similar to those of the training set, the predicted $|\Delta N|$ can be as large as 0.1, and even larger for out-of-sample molecules. The error in the number of electrons leads to the impossibility of reliably computing properties such as other multipole moments, electrostatic potential, or exchange-correlation energy, and thus prompts us to explore and compare different approaches to correct for N both after prediction and during the learning step.

IV. THE NUMBER OF ELECTRONS

In Sec. IV A, we first discuss different ways to correct for the number of electrons given by the approximate density of one molecule either at the decomposition or at the prediction step. Next, in Sec. IV B and Sec. IV C, we propose two modifications for the model to exploit the information about the number of electrons at the learning step.

A. *A posteriori* correction of the predicted densities

By definition, the integral of the exact electron density over all space is the number of electrons N. In our case, when an approximate density is determined by a set of coefficients $\{c_i\}$, it integrates to a value

$$N(\mathbf{c}) = \int \rho(\mathbf{r})\,\mathrm{d}^3\mathbf{r} = \sum_i c_i\, q_i, \tag{17}$$

where $q_i = \int \phi_i(\mathbf{r})\,\mathrm{d}^3\mathbf{r}$ is the charge bearing by the basis function ϕ_i. Even though the loss function of Eq. 10 searches for an approximate electron density being the closest to the reference, it does not contain any explicit constraints, and the fact that we use an incomplete basis set leads to some inaccuracies in the number of electrons as in all other properties. Moreover, the predicted coefficients in the form of Eq. 15 are not constrained either and give a number of electrons close to N only when the prediction errors are small enough.

The correct $N(\mathbf{c})$ for the density fitting has been traditionally achieved by adding a constraint[27] on the number of electrons in the DF loss function (10). Hence, we get another set of decomposition coefficients

$$\mathbf{c}^{\mathrm{DF},N} = \mathbf{c}^{\mathrm{DF}} + \lambda\,\mathbf{O}^{-1}\mathbf{q}, \tag{18}$$

where $\mathbf{c}^{\mathrm{DF}}$ is determined by Eq. 12 and the Lagrange multiplier $\lambda = (N - N(\mathbf{c}^{\mathrm{DF}}))/\mathbf{q}^\intercal\mathbf{O}^{-1}\mathbf{q}$. Even though $q_i \neq 0$ only for spherically symmetric functions, all the basis functions are coupled via the matrix $\mathbf{O}$ and thus participate in the correction of the coefficients.

It is also possible to rewrite Eq. 18 by introducing a nonlinear "operator" $\hat{\mathrm{P}}^{\mathbf{O}}_{\mathrm{N}}$ that acts on any density in the form of a sum of atom-centered contributions (Eq. 9 or 14), giving a new density, which is the closest in O-sense to the "old" one, but integrates to exactly N electrons,

$$\hat{\mathrm{P}}^{\mathbf{O}}_{\mathrm{N}}\,\mathbf{c} = \mathbf{B}\,\mathbf{c} + \mathbf{n} \tag{19}$$

with

$$\mathbf{B} = \mathbf{1} - \frac{\mathbf{O}^{-1}(\mathbf{q}\mathbf{q}^\intercal)}{\mathbf{q}^\intercal\mathbf{O}^{-1}\mathbf{q}}, \quad \mathbf{n} = \frac{N}{\mathbf{q}^\intercal\mathbf{O}^{-1}\mathbf{q}}\cdot\mathbf{O}^{-1}\mathbf{q}. \tag{20}$$

In this way it is possible to correct (to "refit") the coefficients $\mathbf{c}^{\mathrm{ML}}$ predicted by the original — uncorrected — machine learning model and obtain the new coefficients $\mathbf{c}^{\mathrm{ML},N} = \hat{\mathrm{P}}^{\mathbf{O}}_{\mathrm{N}}\,\mathbf{c}^{\mathrm{ML}}$ suitable for computing electrostatic potential, multipole moments, and other extensive properties. Here it is implied that the correction metric O is the same as the ML metric, but this restriction is not compulsory.

For instance, to correct the coefficients predicted for large molecules such as proteins, the straightforward computation of the dot product $\mathbf{O}^{-1}\mathbf{q}$ ($O = S$ or J) is nearly impossible, even though it can be implemented with integral screening and iterative matrix inversion methods. Since we, in principle, can use different metrics for the decomposition, prediction, and correction for the number of electrons, a way to avoid the computational burden of inverting $\mathbf{S}$ or $\mathbf{J}$ is to simply use the unit matrix instead, *i.e.* $\mathbf{O} = \mathbf{1}$, and use an operator $\hat{\mathrm{P}}^{\mathbf{1}}_{\mathrm{N}}$. (This operator only acts on the s-function coefficients.)

Alternatively, an even simpler way to correct the coefficients for the number of electrons is to scale them as

$$\hat{\mathrm{P}}'_{\mathrm{N}}\,\mathbf{c}^{\mathrm{ML}} = \frac{N}{\mathbf{q}^\intercal\mathbf{c}^{\mathrm{ML}}}\cdot\mathbf{c}^{\mathrm{ML}}. \tag{21}$$

This approach is somewhat arbitrary since all the coefficients, even the ones that do not contribute to the number of electrons, are scaled uniformly (one might as well scale only the coefficients for s-functions). In this work, we use refitting with the unit matrix and scaling of the coefficients only to correct the final predictions in order to compare them with a more solid approach of Eq. 19.

B. Constrained learning: from $\mathbf{M_0}$ to $\mathbf{M_L}$

In Sec. IV A we described how to correct for the number of electrons by *a posteriori* modification of the predictions obtained from the original model $\mathbf{M_0}$. These procedures are independent from the regression framework and the number of electrons is never taken into account during the learning step. However, such information could improve the final result. Below, we consider two possibilities to explicitly include the particle number information into the machine learning model.

In the original framework[9,10] $\mathbf{M_0}$, the coefficients for a molecular system m depend on the regression weights $\mathbf{x}$ via the kernel matrix $\mathbf{K}_m$ (Eq. 15), and the working equations following from Eq. 13 (without regularization) are

$$\mathbf{x} = \Big(\sum_{m\in\mathrm{TrS}}\mathbf{K}_m^\intercal\mathbf{O}_m\mathbf{K}_m\Big)^{-1}\Big(\sum_{m\in\mathrm{TrS}}\mathbf{K}_m^\intercal\mathbf{b}_m\Big) \equiv \mathbf{A}^{-1}\mathbf{u}, \tag{22}$$

where $\mathbf{b}_m = \mathbf{O}_m\mathbf{c}_m^{\mathrm{DF}}$ comes from density-fitting coefficients. (If the DF and ML metrics are the same, $\mathbf{b} \equiv \mathbf{w}$.)

Using the Lagrange multipliers method, it is possible to constrain the model to yield N_m electrons for each structure

in the training set (or, generally, in the "constraint set" CS),

$$\Lambda'_{\mathrm{ML}} = \sum_{m\in\mathrm{TrS}} (\rho'_{m,\mathrm{ML}} - \rho_{m,\mathrm{DF}}|\,\hat{\mathrm{O}}\,|\rho'_{m,\mathrm{ML}} - \rho_{m,\mathrm{DF}}) + 2\sum_{m\in\mathrm{CS}} \lambda_m \left(\int \rho'_{m,\mathrm{ML}}\,\mathrm{d}^3\mathbf{r} - N_m \right). \tag{23}$$

The regression weights are

$$\mathbf{x}' = \mathbf{A}^{-1}\mathbf{u} - \sum_{m\in\mathrm{CS}} \mathbf{K}_m^\intercal \mathbf{q}_m \lambda_m \tag{24}$$

and the Lagrange multipliers $\{\lambda_m\}$ are the solution of the linear system

$$\sum_{n\in\mathrm{CS}} (\mathbf{q}_m^\intercal \mathbf{K}_m \mathbf{A}^{-1} \mathbf{K}_n^\intercal \mathbf{q}_n)\lambda_n = \mathbf{q}_m^\intercal \mathbf{K}_m \mathbf{A}^{-1}\mathbf{u} - N_m \quad \forall m \in \mathrm{CS}. \tag{25}$$

(We denote the constrained model $\mathbf{M_L}$ to distinguish it from the original $\mathbf{M_0}$.)

By construction, the regression weights $\mathbf{x}'$ lead to coefficients giving the exact number of electrons for any structure in the training set. Yet, the predicted coefficients for an arbitrary molecule are not under any constraint and should thus be corrected after prediction. A smaller error in the number of electrons is however expected in comparison to the one from the original model.

In principle, we are not restricted to put constraints on the same structures as used for the minimization of Λ'_{ML}. The sums in Eqs. 24 and 25 can be computed over *e.g.* only a part of the training set, the training set and some additional structures, or a completely different set of structures. Despite having a vague physical sense, this flexibility can be exploited for better understanding the model (see Sec. II of the Supplementary Material), for example, by varying the training-set size with constant constraint subset or vice versa.

C. Modification of kernels: from $\mathbf{M_L}$ to $\mathbf{M_K}$

With $\mathbf{M_L}$, the information about the number of electrons is explicitly used in the model, but only for the training-set molecules. Another more consistent possibility is to modify directly the kernel function to ensure the exact number of electrons for any set of coefficients obtained through it. In this way, the molecules in both the training and test sets are treated on the same footing, while the training-set prediction error (*i.e.* ML loss function) is minimized for the corrected densities.

Combining Equations 13, 15, and 19, we get a new model $\mathbf{M_K}$,

$$\Lambda''_{\mathrm{ML}} = \sum_{m\in\mathrm{mol}} (\rho''_{m,\mathrm{ML}} - \rho_{m,\mathrm{DF}}|\,\hat{\mathrm{O}}\,|\rho''_{m,\mathrm{ML}} - \rho_{m,\mathrm{DF}}), \tag{26}$$

where the $\rho''_{m,\mathrm{ML}}$ are determined by coefficients $\mathbf{c}''^{\mathrm{ML}}_m$,

$$\mathbf{c}''^{\mathrm{ML}}_m(\mathbf{x}) = \hat{\mathrm{P}}^{\mathbf{O}}_{\mathrm{N}\,m}\mathbf{c}^{\mathrm{ML}}_m(\mathbf{x}) = \hat{\mathrm{P}}^{\mathbf{O}}_{\mathrm{N}\,m}\mathbf{K}_m\mathbf{x}. \tag{27}$$

The working equations become

$$\mathbf{x}'' = \Big(\sum_{m\in\mathrm{mol}} \mathbf{K}_m^\intercal \tilde{\mathbf{O}}_m \mathbf{K}_m\Big)^{-1} \Big(\sum_{m\in\mathrm{mol}} \mathbf{K}_m^\intercal \tilde{\mathbf{b}}_m\Big) \tag{28}$$

with $\tilde{\mathbf{O}} = \mathbf{B}^\intercal\mathbf{O}\mathbf{B}$ and $\tilde{\mathbf{b}} = \mathbf{B}^\intercal\mathbf{b}$ (since $\mathbf{B}^\intercal\mathbf{n} = 0$). Equation 28 has the same form as Eq. 22, but the original quantum-chemical data $\mathbf{O}$ and $\mathbf{b}$ are transformed by matrix $\mathbf{B}$ defined by Eq. 20.

The final predictions are obtained by using the regression weights $\mathbf{x}''$ in Eq. 27, which is analogous to the prediction according to Eq. 15 followed by the correction.

However, because the matrix $\mathbf{B}$ is idempotent, the modified metric matrix $\tilde{\mathbf{O}} = \mathbf{B}^\intercal\mathbf{O}\mathbf{B}$ by construction has a zero eigenvalue and thus is singular, making the regression problem ill-defined. To get rid of the singularity, we propose to modify the loss function and minimize the prediction error for both the corrected and uncorrected densities simultaneously, by adding to Eq. 26 a small fraction $\alpha \in (0;1)$ of Eq. 13,

$$\Lambda''_{\mathrm{ML}} = \sum_{m\in\mathrm{mol}} \Big(\alpha\cdot(\rho_{m,\mathrm{ML}} - \rho_{m,\mathrm{DF}}|\,\hat{\mathrm{O}}\,|\rho_{m,\mathrm{ML}} - \rho_{m,\mathrm{DF}}) + (1-\alpha)\cdot(\rho''_{m,\mathrm{ML}} - \rho_{m,\mathrm{DF}}|\,\hat{\mathrm{O}}\,|\rho''_{m,\mathrm{ML}} - \rho_{m,\mathrm{DF}})\Big). \tag{29}$$

The working equations are still in the form of (28) with the modified molecular data

$$\tilde{\mathbf{O}} = (1-\alpha)\cdot\mathbf{B}^\intercal\mathbf{O}\mathbf{B} + \alpha\cdot\mathbf{O}, \tag{30}$$

$$\tilde{\mathbf{b}} = (1-\alpha)\cdot\mathbf{B}^\intercal\mathbf{b} + \alpha\cdot\mathbf{b}, \tag{31}$$

we used $\alpha = 10^{-6}$ to make the perturbation small but still have an acceptable condition number of the $\tilde{\mathbf{O}}$ matrix.

V. METRICS, MODELS, AND CORRECTIONS: INFLUENCE ON ESP AND DIPOLE MOMENT

In addition to $\mathbf{M_0}$ (Sec. III B), we computed the predictions for $\mathbf{M_L}$ (model of Eq. 24) and $\mathbf{M_K}$ (model of Eq. 28). However, even though $\mathbf{M_L}$ works as expected on a small set, the linear system of Eq. 25 becomes ill-defined and the constraints cannot be fulfilled on a large enough training set (number of molecules $\approx$ number of reference environments M, see Sec. II of the Supplementary Material for details). For this reason, we have dropped $\mathbf{M_L}$ from the discussion and focus only on $\mathbf{M_0}$ and $\mathbf{M_K}$ hereinafter.

Figure 2 (solid bars) shows the errors with respect to the *ab initio* results for the electrostatic potentials and dipole moments predicted with $\mathbf{M_K}$ and $\mathbf{M_0}$ corrected according to Eq. 19. In comparison with the ESP predicted with the original $\mathbf{M_0}$ model (*i.e.*, 0.8–2.3 kcal/mol in Fig. 1c), the prediction with an *a posteriori* correction leads to errors at least two times smaller (about 0.4 kcal/mol). In contrast and as expected, the DORI similarity measures are not affected by the correction (See Fig. S3 of the Supplementary Material). Including the information about the number of particles into the kernel leads to lower errors in the ESP and dipole moments than those with the *a posteriori* correction alone for both the *JJ* and *SS* combinations.

We also explore the simpler ways to correct the final predictions, *i.e.*, the refitting with the unit matrix and the uniform scaling. Even though the kernel function of $\mathbf{M_K}$ is already

This article appeared in J. Chem. Phys. **155**, 204111 (2021) and may be found at https://aip.scitation.org/doi/10.1063/5.0055393.

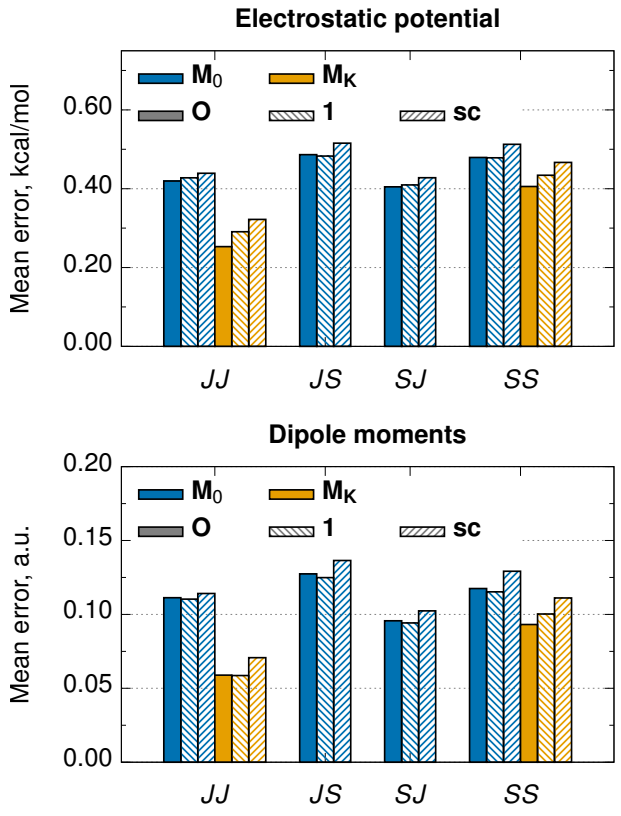

FIG. 2. Comparison of four combinations of metrics (*JJ*, *JS*, *SJ*, and *SS*) and two models ($\mathbf{M_0}$ and $\mathbf{M_K}$): mean errors, computed on the test set, in *(top)* the electrostatic potential on the isosurface $p_0 = 4$ and *(bottom)* dipole moments for predicted densities upon correction for the number of electrons. *Solid bars* (**O**): correction according to Eq. 19 with metric corresponding to the ML metric; *\-filled bars* (**1**): correction using a unit matrix according to Eq. 19; */-filled bars* (**sc**): correction by scaling according to Eq. 21. Errors in the ESP computed on other grids are provided in Table S6.

defined to always lead to the correct N, we also make, for comparison, the final predictions with $\hat{\mathrm{P}}^{\mathbf{1}}_{\mathrm{N}}$ or $\hat{\mathrm{P}}'_{\mathrm{N}}$ instead of $\hat{\mathrm{P}}^{\mathbf{O}}_{\mathrm{N}}$ in Eq. 27. The errors in ESP and dipole moments are shown on Fig. 2, pattern-filled bars. It is notable that correction with the unit metric does not significantly deteriorate the results obtained with the most sophisticated scheme and can thus be used for larger molecules inadequate for $\hat{\mathrm{P}}^{\mathbf{O}}_{\mathrm{N}}$.

We note also that the effect of the metrics is more significant for the learning stage: *S*-learning always gives larger errors than *J*-learning. It is interesting that while *JJ* and *SJ* are nearly the same for ESP, *SJ* works better for the dipole moments. Yet, the $\mathbf{M_K}$ model in combination with the *JJ* metric provides the best overall results on the test set.

VI. EXTRAPOLATION

To validate the results of Sec. V on larger systems, we predict the densities of the same eight oligopeptides taken from the Protein Data Bank as used in our previous work[10] within both the $\mathbf{M_0}$ model with an *a posteriori* correction and the $\mathbf{M_K}$ model. Evaluating the performance of the corrected models on larger molecules is especially relevant because the predictions of the electrostatic potential or multipole moments are not possible with the original $\mathbf{M_0}$ models. The latter indeed yield to large errors in the number of electrons for these oligopeptides (up to two orders of magnitude larger than those for the test set, see Table S9), which makes the computation of any property from the predicted density meaningless.

The errors in the predicted ESP and dipole moments with respect to the *ab initio* ones (normalized by mapping to the $[0,1]$ interval) are shown on Fig. 3a. To our surprise, the metrics *JJ* and *SJ* within $\mathbf{M_0}$, which perform well on the test set, are usually the worst for the oligopeptides. Moreover, the least physically sound correction scheme — scaling — performs generally better than the sophisticated $\hat{\mathrm{P}}^{\mathbf{O}}_{\mathrm{N}}$, suggesting an error cancellation.

Within this context, it is important to stress that our original training set is based only on the side-chain–side-chain dimer subset of BFDb with no explicit representation of peptide bonds. For this reason, the highest absolute errors in the predicted densities were shown to be mostly localized on the oligopeptide backbones.[10] In order to distinguish the effect of increasing the system size from the one originating from the lack of peptide backbones in the training set, the peptide bonds were "cut" and the amino and carboxyl groups were replaced with hydrogen atoms. Already within the non-corrected $\mathbf{M_0}$ model, the average errors in the number of electrons for these "no-backbone" systems are an order of magnitude smaller than those for the original structures (see Table S9). For the corrected models, the normalized errors in the predicted properties are shown on Fig. 3b with the absolute errors shown in Fig. S4b. On average, the absolute errors are 3–6 times lower than those for the original oligopeptides, which confirms the significant perturbation associated with the peptide bonds, while comparing the different models and corrections. This problem, which is not the topic of this work, could be easily addressed by extending the training set. The error spread (Fig. S4b) also decreases, *e.g.*, the ESP errors for 3OW9 lie between 1.5 and 6.6 kcal/mol, whereas for its no-backbone version the interval is $(0.6, 1.6)$ kcal/mol. Overall, all the models and metrics perform very similarly and lead to fairly impressive predictions. Akin to the test set, the $\mathbf{M_K}$, *JJ* combination with any correction scheme offers the best compromise as it leads to the most accurate predictions for most oligopeptides. Similarly, the performance of the *a posteriori* corrected $\mathbf{M_0}$, *JS* models, which was slightly inferior for the test set, is also less robust for the oligopeptide set (for additional comparisons on the oligopeptide set, refer to Figure S5 and Table S9 in the Supplementary Material).

VII. CONCLUSIONS

The analysis of the interplay between deductive reasoning based on quantum-chemical knowledge and the inductive nature of statistical learning is a fundamental issue to further

This article appeared in J. Chem. Phys. **155**, 204111 (2021) and may be found at https://aip.scitation.org/doi/10.1063/5.0055393.

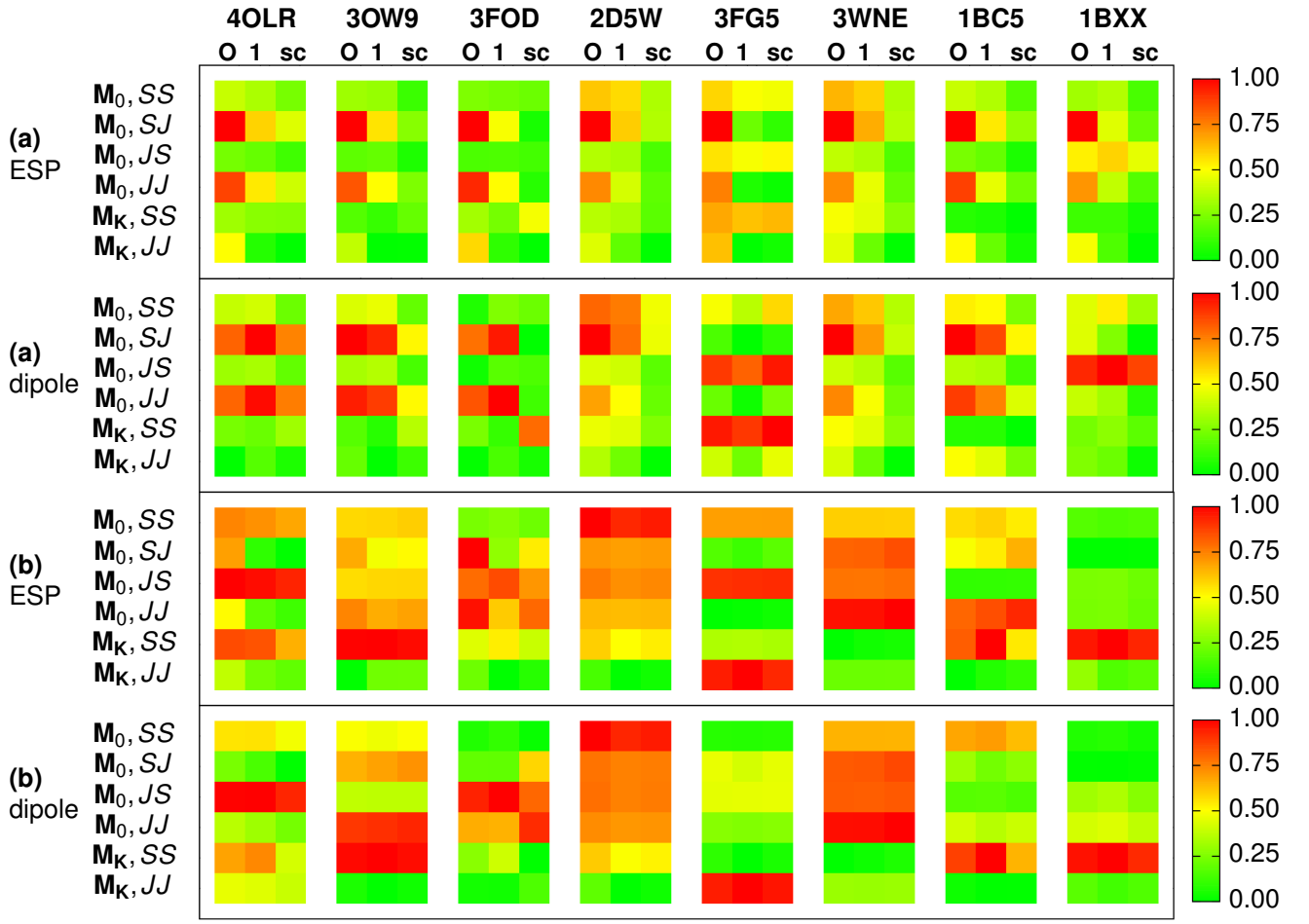

FIG. 3. Comparison of the four combinations of metrics (*JJ*, *JS*, *SJ*, and *SS*), two models ($\mathbf{M_0}$ and $\mathbf{M_K}$), and three ways to correct for the number of electrons (**O**, **1**, and **sc**) by prediction for eight oligopeptides (labeled by PDB ID). Each square represents an error in the predicted electrostatic potential on the isosurface $p_0 = 0.125$ or dipole moments of *(a)* oligopeptides and *(b)* "no-backbone" oligopeptides. For the sake of clarity, the errors for each structure and property (*i.e.* within each 3×6 rectangle) are mapped to the $[0,1]$ interval, *i.e.* $\mathrm{val} \mapsto (\mathrm{val} - \mathrm{val}_{\mathrm{min}})/(\mathrm{val}_{\mathrm{max}} - \mathrm{val}_{\mathrm{min}})$.

improve quantum machine learning models. In this work, we analyze the effects of varying the quantum-chemical metrics used for the decomposition and regression of the molecular electron density. We find that the machine learning loss function is more affected by the choice of metric than the loss function associated with decomposition but overall, the *JJ*-scheme shows the best performance. Yet, the learning exercise is equally difficult regardless of the metric used to decompose the density.

Importantly, imposing the correct number of electrons appears crucial to accurately predict extensive properties such as the ESP and multipole moments. Correcting the predictions *a posteriori* for the number of electrons makes the accuracy of the $\mathbf{M_0}$ model largely independent from the choice of the quantum-chemical metric. This result is especially important for periodic systems or for situations where the charge density (or another density-like object) can be obtained only on a real-space grid for which the Coulomb metric is ill-defined and the overlap has to be used. As a step forward, we propose the $\mathbf{M_K}$ model, in which the kernels explicitly include the information about the number of electrons. While both *a posteriori* correction and kernel modification increase slightly the computational complexity on the prediction step, it is always possible to apply other corrections (such as the unit-matrix correction) when extrapolating on larger chemical systems.

Overall, this work demonstrates that choosing a proper quantum-chemical metric to optimize ML models is important and that this is especially true if the model is not built to encode all the proper fundamental physical constraints.[42]

SUPPLEMENTARY MATERIAL

See the Supplementary Material for the learning curves, discussion of the $\mathbf{M_L}$ model, additional numerical data, and statistical analysis.

ACKNOWLEDGMENTS

The authors thank Andrea Grisafi, David M. Wilkins, and Michele Ceriotti for sharing the code[14] to construct the tensorial SOAP kernels. A.F. acknowledges financial support from the National Centre of Competence in Research (NCCR) "Materials' Revolution: Computational Design and Discovery of Novel Materials (MARVEL)" of the Swiss National Science Foundation (SNSF). K.B. was supported by the European Research Council (ERC, grant agreement no 817977).

DATA AVAILABILITY

The data and the model that support the findings of this study are freely available on the Materials Cloud at `https://doi.org/10.24435/materialscloud:d8-0h`.

[1] T. S. Koritsanszky and P. Coppens, Chem. Rev. **101**, 1583 (2001).
[2] J. C. Meyer, S. Kurasch, H. J. Park, V. Skakalova, D. Künzel, A. Groß, A. Chuvilin, G. Algara-Siller, S. Roth, T. Iwasaki, U. Starke, J. H. Smet, and U. Kaiser, Nat. Mater. **10**, 209 (2011).
[3] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[4] J. M. Alred, K. V. Bets, Y. Xie, and B. I. Yakobson, Compos. Sci. Technol. **166**, 3 (2018).
[5] A. T. Fowler, C. J. Pickard, and J. A. Elliott, J. Phys. Mater. **2**, 034001 (2019).
[6] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad, npj Comput. Mater. **5**, 22 (2019).
[7] P. B. Jørgensen and A. Bhowmik, "DeepDFT: Neural message passing network for accurate charge density prediction," (2020), arXiv:2011.03346 [physics.comp-ph].
[8] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, Nat. Commun. **8**, 872 (2017).
[9] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, ACS Cent. Sci. **5**, 57 (2018).
[10] A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Corminboeuf, Chem. Sci. **10**, 9424 (2019).
[11] A. Fabrizio, K. Briling, A. Grisafi, and C. Corminboeuf, CHIMIA **74**, 232 (2020).
[12] A. Fabrizio, K. R. Briling, D. D. Girardier, and C. Corminboeuf, J. Chem. Phys. **153**, 204111 (2020).
[13] A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, Phys. Rev. Lett. **120**, 036002 (2018).
[14] `https://github.com/dilkins/TENSOAP/`.
[15] M. D. Newton, J. Chem. Phys. **51**, 3917 (1969).
[16] E. J. Baerends, D. E. Ellis, and P. Ros, Chem. Phys. **2**, 41 (1973).
[17] H. Sambe and R. H. Felton, J. Chem. Phys. **62**, 1122 (1975).
[18] C. Van Alsenoy, J. Comput. Chem. **9**, 620 (1988).
[19] O. Vahtras, J. Almlöf, and M. W. Feyereisen, Chem. Phys. Lett. **213**, 514 (1993).
[20] J. W. Mintmire and B. I. Dunlap, Phys. Rev. A **25**, 88 (1982).
[21] C.-K. Skylaris, L. Gagliardi, N. C. Handy, A. G. Ioannou, S. Spencer, and A. Willetts, J. Mol. Struct. THEOCHEM **501-502**, 229 (2000).
[22] Y. Jung, A. Sodt, P. M. W. Gill, and M. Head-Gordon, Proc. Natl. Acad. Sci. U.S.A. **102**, 6692 (2005).
[23] J. L. Whitten, J. Chem. Phys. **58**, 4496 (1973).
[24] B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, J. Chem. Phys. **71**, 3396 (1979).
[25] B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, J. Chem. Phys. **71**, 4993 (1979).
[26] G. G. Hall and C. M. Smith, Int. J. Quantum Chem. **25**, 881 (1984).
[27] C. M. Smith and G. G. Hall, Theor. Chim. Acta **69**, 63 (1986).
[28] K. Eichkorn, O. Treutler, H. Öhm, M. Häser, and R. Ahlrichs, Chem. Phys. Lett. **240**, 283 (1995).
[29] P. M. W. Gill, B. G. Johnson, J. A. Pople, and S. W. Taylor, J. Chem. Phys. **96**, 7178 (1992).
[30] S. Reine, E. Tellgren, A. Krapp, T. Kjærgaard, T. Helgaker, B. Jansik, S. Høst, and P. Salek, J. Chem. Phys. **129**, 104101 (2008).
[31] P. de Silva and C. Corminboeuf, J. Chem. Theory Comput. **10**, 3745 (2014).
[32] L. A. Burns, J. C. Faver, Z. Zheng, M. S. Marshall, D. G. A. Smith, K. Vanommeslaeghe, A. D. MacKerell, K. M. Merz, and C. D. Sherrill, J. Chem. Phys. **147**, 161727 (2017).
[33] Q. Sun, J. Comput. Chem. **36**, 1664 (2015).
[34] Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, and G. K.-L. Chan, Wiley Interdiscip. Rev. Comput. Mol. Sci. **8**, e1340 (2017).
[35] J.-D. Chai and M. Head-Gordon, Phys. Chem. Chem. Phys. **10**, 6615 (2008).
[36] T. H. Dunning, J. Chem. Phys. **90**, 1007 (1989).
[37] F. Weigend, Phys. Chem. Chem. Phys. **4**, 4285 (2002).
[38] D. N. Laikov, J. Chem. Phys. **135**, 134120 (2011).
[39] V. I. Lebedev, Zh. Vychisl. Mat. Mat. Fiz. **16**, 293 (1976).
[40] V. I. Lebedev and D. N. Laikov, Russ. Acad. Sci. Dokl. Math. **59**, 477 (1999).
[41] P. D. Walker and P. G. Mezey, J. Am. Chem. Soc. **116**, 12022 (1994).
[42] While the framework can, in principle, accommodate any molecular property as constraint, the computational advantages of machine learning are leveraged only when using readily obtainable quantities such as the number of electrons, which do not require any quantum-chemical computation.