Supplementary Information

Electron density learning of non-covalent systems

Alberto Fabrizio,[a,b] Andrea Grisafi,[c,b] Benjamin Meyer,[a,b] Michele Ceriotti,[c,b] and Clemence Corminboeuf*[a,b]

[a] *Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.*

[b] *National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

[c] *Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland*

E-mail: clemence.corminboeuf@epfl.ch

1 Density decomposition using a density-fitting basis

Within the density-fitting framework two-center densities are approximated using an auxiliary-basis ($\phi_P^{RI}(\boldsymbol{r})$) as

$$\rho_{pq}(\boldsymbol{r}) = \chi_p(\boldsymbol{r})\chi_q(\boldsymbol{r}) = \sum_P^{N_{aux}} d_P^{pq}\, \phi_P^{RI}(\boldsymbol{r}) \tag{S1}$$

where $\chi_q(\boldsymbol{r})$ are the atomic orbitals and d_P^{pq} are the RI-expansion coefficients defined as

$$d_P^{ab} = \sum_Q (ab|Q)[\boldsymbol{J}^{-1}]_{QP} \tag{S2}$$

Equation S2 relies upon the following three-center two electron integrals

$$(ab|Q) = \int d\boldsymbol{r_1} \int d\boldsymbol{r_2} \chi_a(\boldsymbol{r_1})\chi_b(\boldsymbol{r_1})\frac{1}{r_{12}}\phi_Q^{RI}(\boldsymbol{r_2}) \tag{S3}$$

while $[\boldsymbol{J}^{-1}]_{QP}$ is defined as:

$$[\boldsymbol{J}]_{QP} = \int d\boldsymbol{r_1} \int d\boldsymbol{r_2} \phi_Q^{RI}(\boldsymbol{r_2})\frac{1}{r_{12}}\phi_P^{RI}(\boldsymbol{r_2}) \tag{S4}$$

Then, starting from the definition of the electron density and using Eq. S1, we can write:

$$\rho(\boldsymbol{r}) = \sum_{ab}^{N_{AO}} D_{ab}\, \chi_a(\boldsymbol{r})\chi_b(\boldsymbol{r}) \tag{S5}$$

$$\rho(\boldsymbol{r}) = \sum_{ab}^{N_{AO}} D_{ab}\, \chi_a(\boldsymbol{r})\chi_b(\boldsymbol{r}) = \sum_{ab}^{N_{AO}} D_{ab}\, (\sum_{P}^{N_{aux}} d_P^{ab}\, \phi_P^{RI}(\boldsymbol{r})) = \tag{S6}$$

$$= \sum_{P}^{N_{aux}} (\sum_{ab}^{N_{AO}} D_{ab}\, d_P^{ab}\,)\phi_P^{RI}(\boldsymbol{r})) = \tag{S7}$$

$$= \sum_{P}^{N_{aux}} C_P \phi_P^{RI}(\boldsymbol{r})) \tag{S8}$$

where D_{ab} is the 1-electron reduced density matrix.

2 Symmetry-Adapted Machine Learning Model

The core idea of our density-learning framework consists in representing ρ in a non-orthogonal basis set $\{\phi_k\}$ made of spherical harmonics and radial functions centered on each atom of the molecule, i.e.,

$$\rho(\boldsymbol{r}) = \sum_{i\in\{\text{atoms}\}} \sum_k c_k^i \phi_k(\boldsymbol{r} - \boldsymbol{r}_i) = \sum_{i\in\{\text{atoms}\}} \sum_{nlm} c_{nlm}^i R_n(|\boldsymbol{r} - \boldsymbol{r}_i|) Y_{lm}\left(\widehat{\frac{\boldsymbol{r} - \boldsymbol{r}_i}{|\boldsymbol{r} - \boldsymbol{r}_i|}}\right) . \tag{S9}$$

In a nutshell, one can ask to predict the expansion coefficients $\{c_k^i\}$ by training a symmetry-adapted Gaussian process regression model on a set of N reference densities $\{\rho\}_N$ previously computed[1]. To do so, symmetry adapted kernels[2] are constructed from the atomic coordinates of the molecules in order to represent the statistical correlations of the expansion coefficients, interpreted as spherical tensors of a given angular momentum order l, that is $k_{mm'}^l(i,j) \equiv \langle c_{nlm}^i, c_{nlm'}^j \rangle$. Once these kernel functions are computed for each angular momentum value included in the expansion of $\rho(\boldsymbol{r})$, the learning can be carried out and the spherical harmonic components of the density of any given new molecule can be predicted. In compact notation, the vector $\boldsymbol{c}_{\mathcal{A}}$ that contains the predicted expansion coefficients of a molecule $\mathcal{A}$ is obtained as follows

$$\boldsymbol{c}_{\mathcal{A}} = \boldsymbol{k}_{\mathcal{A}M}\boldsymbol{x}_M = \boldsymbol{k}_{\mathcal{A}M}\left(\boldsymbol{K}_{NM}^T \boldsymbol{S}_{NN} \boldsymbol{K}_{NM} + \eta \boldsymbol{K}_{MM}\right)^{-1} \boldsymbol{K}_{NM}^T \boldsymbol{w}_N\ . \tag{S10}$$

Here, the vector $\boldsymbol{w}_N$ contains the projections $\langle\phi|\rho\rangle$ of the N reference densities on the basis functions and $\boldsymbol{S}_{NN}$ is a block-diagonal matrix containing the spatial overlap $\langle\phi|\phi'\rangle$ between the basis functions of each reference molecule. The value M identifies the number of atomic environments which can be suitably selected to remap the density learning problem in a representative subset of the total number of atoms belonging to the ensemble of the N training molecules. The rectangular matrix $\boldsymbol{K}_{NM}$ contains the symmetry-adapted kernels coupling the atoms of the training set with the ones of the sparse representative set, while the square matrix $\boldsymbol{K}_{MM}$ only couples the atoms within the sparse set. The parameter η needs to be empirically adjusted in order to reach the best compromise between fitting accuracy and smoothness of the model. Note that the dimensionality of the final problem to be solved is defined by M times the number of basis functions that are associated with each selected atom. The outcome of the learning procedure is given by the vector $\boldsymbol{x}_M$ which contains the set of regression weights. These weights can be finally used in combination with the kernel matrix $\boldsymbol{k}_{\mathcal{A}M}$ to obtain the predicted expansion coefficients for the molecule $\mathcal{A}$.

It is worth noticing that if the basis functions were made orthogonal, then $\boldsymbol{S}_{NN} = \mathbf{1}_{NN}$, and the problem of learning $\rho(\boldsymbol{r})$ can be conveniently recast into the problem of learning each set of independent orthogonal projections. This is indeed exactly what is done in Ref. 3 by expanding the electron density in a plane-wave basis set and learning each Fourier component of $\rho(\boldsymbol{r})$ separately. From a computational point of view, this framework is undoubtedly more convenient than having to deal with all the set of density projections at once. However, we believe that the resulting increase in complexity due to the use of a non-orthogonal basis is greatly paid back by the local nature of our regression model, which allows for highly transferable predictions across very diverse chemical and conformational subspaces.

Kernel functions that are covariant in $\mathcal{O}(3)$ have been constructed for each of these angular momentum values, according to the l-SOAP formalism discussed in Refs. 2,4. In particular, an environment cutoff of 4.0 Å and a Gaussian smearing of 0.3 Å have been used to define the SOAP-density around each atom of the dataset. The resulting symmetry-adapted SOAP descriptors are sparsified with farthest point sampling down to 1000 components. A non-linearity degree of order $\zeta = 2$ has been used to enhance the regression power. The representative set of atoms, needed to contract the dimensionality of the density-learning problem, has been defined by selecting, with farthest point sampling, the most diverse M atomic environments according to the scalar 0-SOAP metric. This procedure accounts for a sparsification of the training set of more than an order of magnitude, going from about 40’000 atomic environments to 1000 environments at the most. When solving the regression problem of Eq. (S10), a regularization parameter $\eta = 10^{-5}$ has been found to yield optimal predictions.

2.1 Density Baseline

The direct application of the formula of Eq. (S10) to predict the all-electron density of a molecule would imply that a great portion of the regression effort is spent on capturing the density contributions associated with the core-electrons. Given that these contributions are merely constants across the dataset, it is convenient to apply a baseline to the reference densities and let the regression focuses on the sole chemically driven fluctuations of $\rho(\boldsymbol{r})$. This baseline is chosen such as considering, for each radial function n, the average of the spherical components of the density across the dataset, i.e., $\bar{c}_{n00}$. For each molecule $\mathcal{A}$, these average components build up an effective density field $\bar{\rho}(\boldsymbol{r})$ given by the superposition of spherically symmetric contributions. The projections of $\bar{\rho}(\boldsymbol{r})$ onto the atomic basis functions can be eventually obtained by considering $\bar{\boldsymbol{w}}_{\mathcal{A}} = \boldsymbol{S}_{\mathcal{A}}\bar{\boldsymbol{c}}_{\mathcal{A}}$, with $\bar{\boldsymbol{c}}_{\mathcal{A}}$ a sparse vector representation of the expansion coefficients which include the set of $\bar{c}_{n00}$. At this point, a much more stable regression can be obtained by including the set of baselined projections $\boldsymbol{w}_N - \bar{\boldsymbol{w}}_N$ in Eq. (S10), and adding the average density components $\bar{c}_{n00}$ back on the predicted expansion coefficients.

3 Computational Details

All DFT computations have been performed with ORCA[5] and a locally modified version of PySCF.[6] In addition, the modified PySCF has been used for the density decomposition and to compute density-fitting integrals through an interface with libcint.[7] All the figures of densities, DORI mappings and electrostatic potentials have been generated using Paraview 5.6. The structures of the polypeptides were taken from the crystallographic data of the Protein DataBank (PDB).[8] Hydrogen positions of the polypeptides were let relax at HF-3c[9] as implemented in ORCA.

4 Comparison between SA-GPR and promolecular densities

The superposition of atomic densities has been successfully used with the NCI analysis to qualitatively characterize non-covalent patterns in large molecular systems (such as DNA and proteins) at a minimal computational cost. While, by construction, promolecular densities do not include the effects of density relaxation, they are able to describe sufficiently well the low-density, low-gradient regions typical of the non-covalent regions.[10] In Figure 1, the example of the enkaphalin (PBD ID: 4OLR) polypeptide is used to compare the density error with respect to ωB97X-D/cc-pVQZ between the SA-GPR and the promolecular density.

As reported in Figure 1, the mean absolute percentage error of the promolecular density (22.97%) is about 17 times larger than with SA-GPR (1.33%), for a similar computational cost once the machine-learning model is trained. The consequences are well illustrated by the DORI analysis of the non-covalent interactions in enkephalin. In particular, the use of promolecular density results in spuriously fragmented DORI domains in the covalent region of the molecule. In addition, quantitative measures of the DORI error show large deviations in the non-covalent regions (e.g., both in the basin volumes (Err. Vol.) and integration (Err. Int.)). As example, the Err. Vol. in the

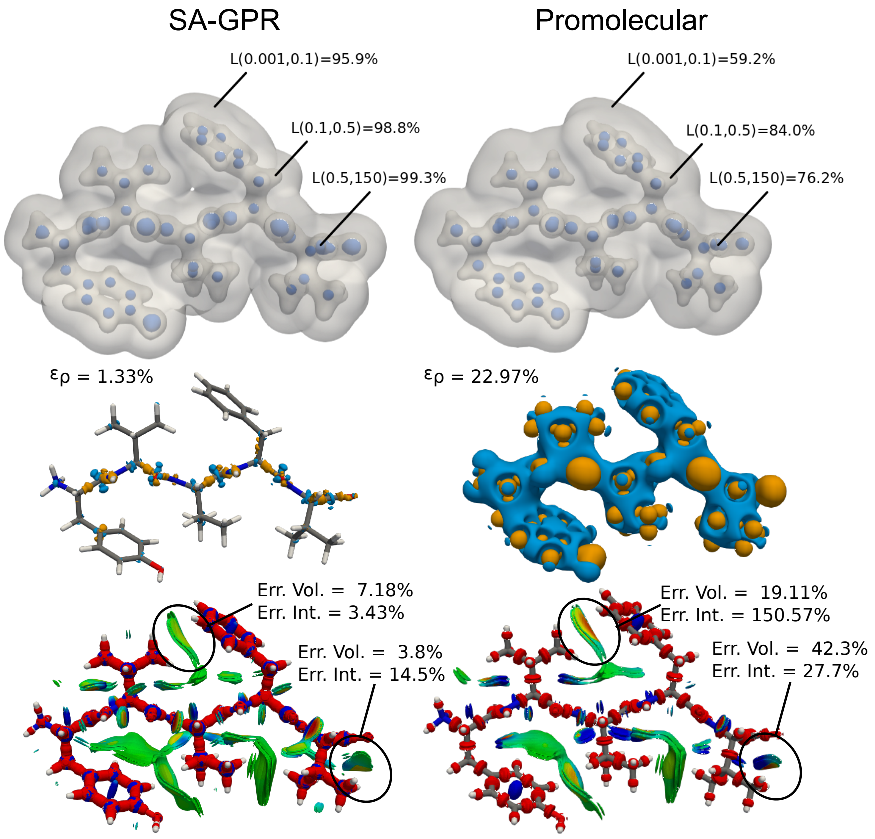

Figure S1: (*top*) Predicted and promolecular density of enkephalin at three isovalues (0.5, 0.1, and 0.001 e^- Bohr^{-3}) enclosing the domain of integration for the Walker-Mezey L(a,a') similarity index with respect to *ab-initio* density. (*middle*) Charge density difference with respect to the ab-initio electron density (isovalues ±0.01 e^- Bohr^{-3}). The mean absolute percentage error is also reported. (*bottom*) DORI map of enkephalin (DORI isovalue: 0.9) colored by $sgn(\lambda_2)\rho(\boldsymbol{r})$ in the range from -0.02 a.u. (red) to 0.02 a.u. (blue).

non-covalent DORI domain between the phenilalanine and the leucine side-chain (see Figure) using the promolecular density are nearly 3 times larger than with SA-GPR. DORI integrals shows a dramatic overestimation of the density in the basin with the promolecular approach (150% of the ab-initio value), compared to only 3.43% for the predicted DORI. Overall, for a similar gain in computational cost, SA-GPR captures the distortion of the atom-in-molecule densities much more accurately than the promolecular approach.

5 Volume and Integral of the DORI basins

The error on DORI can be quantified through the analysis of the volume of the basins and through the integration of the electron density in the volume, which they enclose. Table S1 reports these quantities for the non-covalent basins of the three representative dimers reported in the manuscript. The absolute percentage deviation is reported with respect to the *ab-initio* (cc-pVQZ-RI) value.

Table S1: Volume ($\mathrm{DORI}_{\mathrm{vol}}$) and integrals ($\mathrm{DORI}_{\mathrm{int}}$) of the electron density within DORI [0.9]-domains in a.u. computed from predicted and *ab-initio* electron densities.

	Ab-Initio (cc-pVQZ-RI)		SA-GPR Predicted		Deviation (absolute deviation %)	
	$\mathrm{DORI}_{\mathrm{vol}}$	$\mathrm{DORI}_{\mathrm{int}}$	$\mathrm{DORI}_{\mathrm{vol}}$	$\mathrm{DORI}_{\mathrm{int}}$	$\mathrm{DORI}_{\mathrm{vol}}$	$\mathrm{DORI}_{\mathrm{int}}$
London Dispersion	5.292	0.0259	5.393	0.0261	-0.101 (1.9%)	-0.0002 (0.77%)
Electrostatics	0.947	0.0318	0.938	0.0320	0.009 (0.95%)	-0.0002 (0.63%)
Mixed Regime	9.270	0.0551	9.114	0.0564	0.156 (1.68%)	-0.0013 (2.36%)

6 Comparison between cc-pVXZ and cc-pVXZ-RI

With respect to standard quantum chemical basis sets, density-fitting auxiliary basis (RI-basis) are characterized by higher angular momenta (mainly $\mathrm{l}_{\mathrm{max}} + 1$), slightly more diffuse functions and are, in general, not contracted.[11–13] To further analyze which of these properties has the strongest impact on the accuracy of the density decomposition, we computed the decomposition error on the water molecule with the cc-pVXZ series (Figure 2).

As mentioned in the main text, the addition of diffuse functions (aug-cc-pVXZ series) is not discriminant for the performance and actually adds instabilities in the overlap matrix inversion. Starting with s functions only for cc-pVQZ [cc-pVQZ(s-only)] and gradually adding back all the other angular momenta shows that the overall performance of a basis is already largely determined by l=0, as already discussed in our previous work. The contribution of higher angular momenta is determinant for capturing the deviation from the spherical/isotropic behavior of the density, but is not the main source of difference with the RI-basis. Instead, decontraction is the key element, which alone leads to an error reduction of an order of magnitude. This result is not surprising considering that decontraction of the original basis set (at least in the core region), prior to exponent optimization,

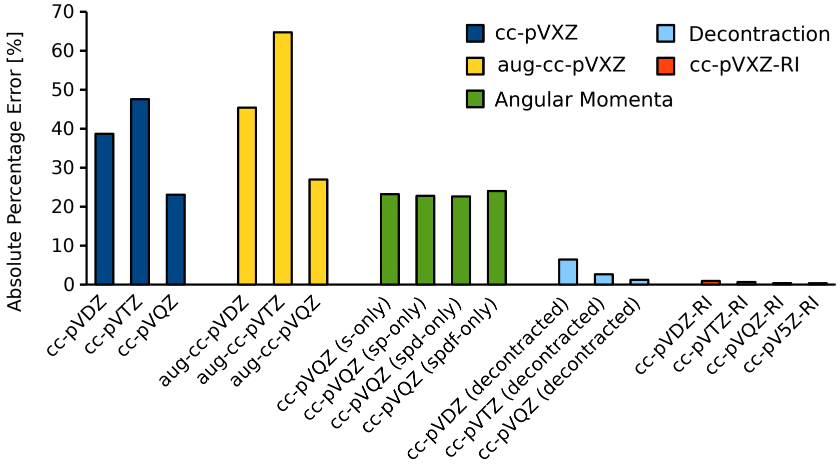

Figure S2: Decomposition error of the electron density for a single water molecule: evolution of the absolute percentage error depending on the choice of basis set. In particular adding diffuse functions, gradually including higher angular momenta, upon decontraction and using resolution of identity auxiliary basis.

is commonly used as a first step for the generation of density-fitting auxiliary basis sets.[13] Beside this aspect, decontraction results in a larger and more flexible basis, more adapted to capture the feature of the density field.

References

[S1] A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf and M. Ceriotti, *ACS Centr. Sci.*, 2019, **5**, 57–64.

[S2] A. Grisafi, D. M. Wilkins, G. Csányi and M. Ceriotti, *Phys. Rev. Lett.*, 2018, **120**, 036002.

[S3] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K.-R. Müller, *Nat. Commun.*, 2017, **8**, 872.

[S4] A. Grisafi, D. M. Wilkins, M. J. Willatt and M. Ceriotti, *arXiv:1904.01623*, 2019.

[S5] F. Neese, *Wiley Interdiscip. Rev. Comput. Mol. Sci*, 2012, **2**, 73–78.

[S6] 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.*, 2018, **8**, e1340.

[S7] Q. Sun, *J. Comput. Chem.*, 2015, **36**, 1664–1671.

[S8] H. M. Berman, *Nucleic Acids Res.*, 2000, **28**, 235–242.

[S9] R. Sure and S. Grimme, *J. Comput. Chem.*, 2013, **34**, 1672–1685.

[S10] E. R. Johnson, S. Keinan, P. Mori-Sánchez, J. Contreras-García, A. J. Cohen and W. Yang, *J. Am. Chem. Soc*, 2010, **132**, 6498–6506.

[S11] F. Weigend, M. Häser, H. Patzelt and R. Ahlrichs, *Chem. Phys. Lett.*, 1998, **294**, 143–152.

[S12] F. Weigend, A. Köhn and C. Hättig, *J. Chem. Phys.*, 2002, **116**, 3175–3183.

[S13] G. L. Stoychev, A. A. Auer and F. Neese, *J. Chem. Theory Comput.*, 2017, **13**, 554–562.