SUPPLEMENTARY INFORMATION

# SPA$^{H}$M: the Spectrum of Approximated Hamiltonian Matrices representations

Alberto Fabrizio,[1,2] Ksenia R. Briling,[1] 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: 16 March 2022)

## CONTENTS

**I. Influence of the basis set and the potential on SPA$^{H}$M** S2

**II. Metallic complexes, large molecules, and conformational diversity** S4

**III. Numerical Data** S5

**References** S8

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

## I. INFLUENCE OF THE BASIS SET AND THE POTENTIAL ON SPA$^{H}$M

The size of the electronic Hamiltonian in the atomic orbitals basis depends on the size of the chosen basis set. Nonetheless, the number of occupied molecular orbitals, which fixes SPA$^{H}$M, only depends on the total number of electrons. In the main text, we have shown the learning curves of the different SPA$^{H}$M representations using a minimal basis set (MINAO).[S1] Figure S1 demonstrates that the overall accuracy of SPA$^{H}$M is largely independent of the choice of the atomic basis. The basis sets reported in the Figure are dramatically different in size and serve distinct conceptual purposes (*e.g.* converging correlation, capturing polarization effects, describing diffuse electron clouds *etc.*). At full training, no significant difference is observable across the entire basis set spectrum.

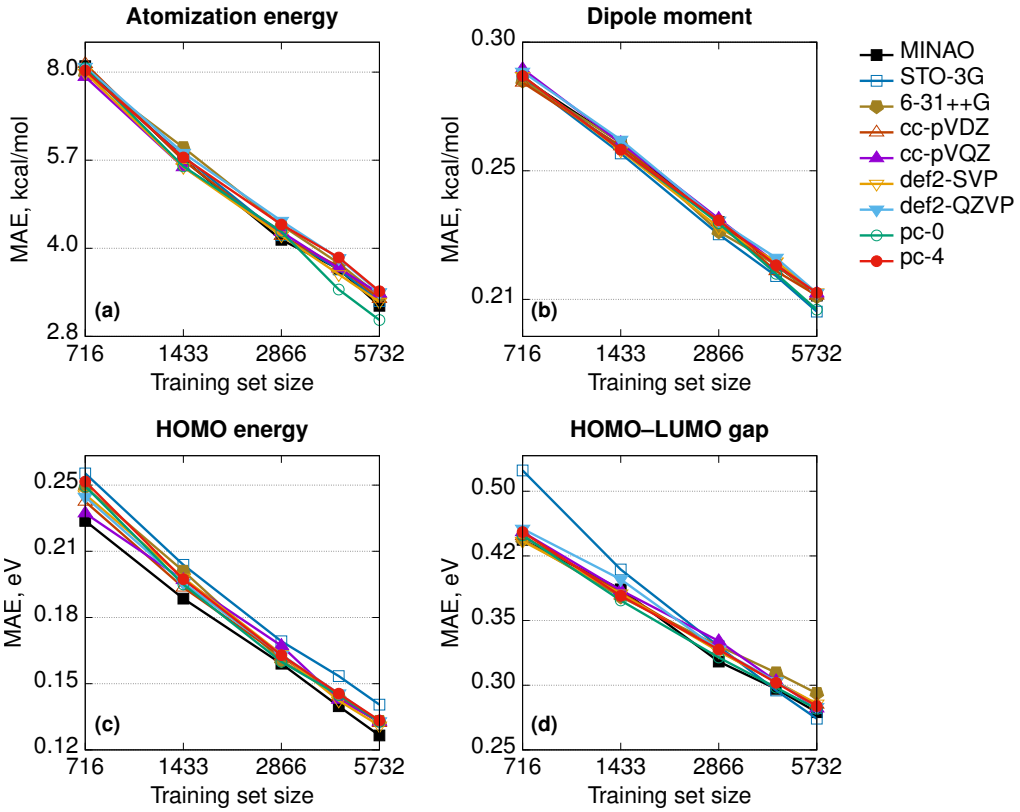


FIG. S1. Learning curves for SPA$^{H}$M based on the LB Hamiltonian using different basis sets.

All the quantum chemical guesses used in SPA$^{H}$M pass through the construction of an approximate Hamiltonian, except for SAD that gives directly a density matrix.[S2,S3] SPA$^{H}$M-SAD is, therefore, the only representation that requires the user to choose a potential. In Figure S2, we show that although different potentials have a larger impact on the quality of the regression than the basis set, the changes are never sufficient to qualitatively change the behavior of the representations.

Interestingly, the simplest of the potentials, Hartree–Fock, leads to most accurate learning both as SPA$^{H}$M-SAD or as converged Fock matrix. This result, particularly evident in the case of the atomization energies, corroborates the conclusion drawn in the main text: adding more physics is not always met with favorable response by the machine learning algorithm. The worsening of the learning curves from HF to density functionals suggests that the introduction of an (approximated) electron correlation potential has the same effect of sparsification of the data in the representation space as observed for the SCF procedure.

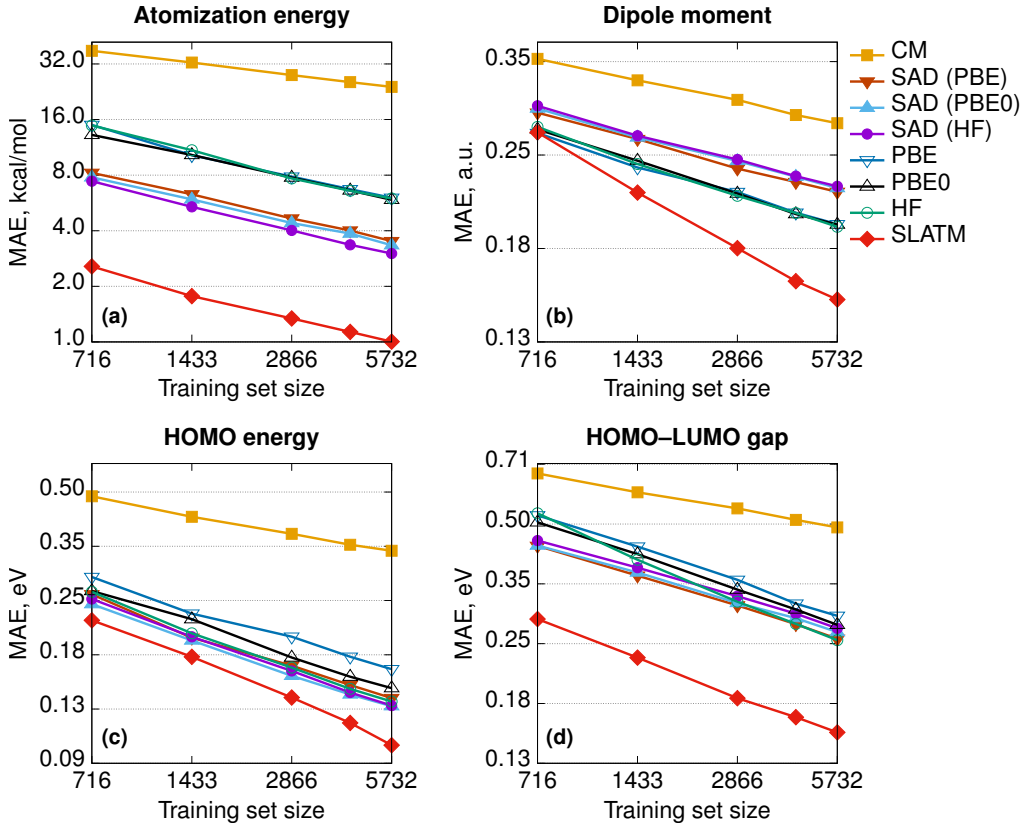


FIG. S2. Learning curves for SPA$^{H}$M based on the SAD guess and converged Fock matrices with different potentials. The learning curves for CM and SLATM are shown for comparison.

## II. METALLIC COMPLEXES, LARGE MOLECULES, AND CONFORMATIONAL DIVERSITY

Relying directly upon the information contained in the electronic Hamiltonian, the SPA$^{H}$M philosophy naturally satisfies the injectivity criterion of quantum machine learning representations. In this work, the practical realization of the SPA$^{H}$M concept takes the form of an eigenvalue spectrum, whose advantage is being naturally invariant under the basic symmetries of physics and well-defined for a molecular representation. Nonetheless, larger molecules and transition metal complexes are usually characterized by a more dense spectrum than the small molecules included in the QM7 and L11 databases.

To test the reliability of eigenvalue SPA$^{H}$M representations on more complex molecules, we report in Figure S3 the accuracy of the SPA$^{H}$M-LB (LB), eigenvalue Coulomb Matrix (CM), and SLATM on a previously published database containing 1473 nickel complexes of varying size (from 22 atoms to 160 atoms).[S4] The organometallic database is freely available at DOI:10.24435/materialscloud:fz-sw. In addition, we test the applicability of eigenvalue SPA$^{H}$M to learn the atomization energy among conformers (*i.e.* constitutional and structural isomers and stereoisomers) on the QM7-X database[S5] (Figure S4).

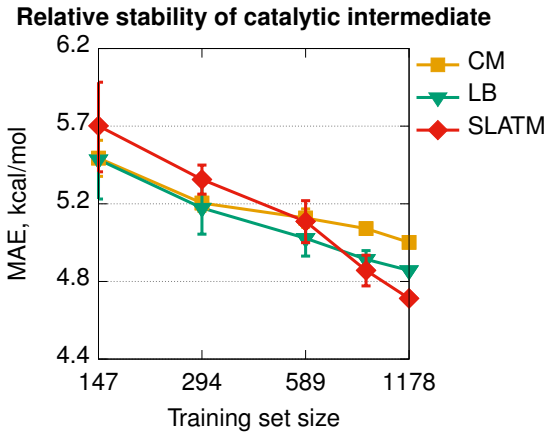


FIG. S3. Learning curves for SPA$^{H}$M based on the LB Hamiltonian, eigenvalue Coulomb Matrix (CM), and SLATM on a database of 1473 nickel complexes. Regression target: energy descriptor [kcal/mol].

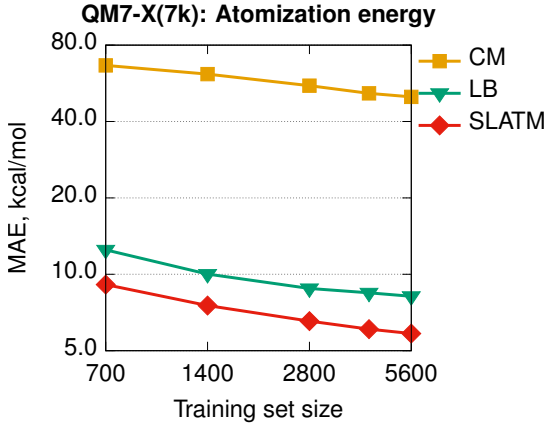


FIG. S4. Learning curves for SPA$^{H}$M based on the LB Hamiltonian, eigenvalue Coulomb Matrix (CM), and SLATM on the QM7-X database.[S5]

In both databases, the performance of SPA$^{H}$M relative to existing representations is similar to the one reported in the main text for QM7, particularly at full training set.

## III. NUMERICAL DATA

TABLE S1. Numerical values for the learning curves (Figure 1)

| Training | **Atomization energy MAE (SD), kcal/mol** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| set size | CM | $H_{\text{core}}$ | GWH | Hückel | SAP | SAD | LB | SLATM |
| 716 | 37.7 (0.7) | 19.0 (0.5) | 18.0 (0.4) | 10 (1) | 11.9 (0.7) | 7.8 (0.4) | 8.2 (0.4) | 2.56 (0.03) |
| 1433 | 33 (1) | 16.2 (0.4) | 13.9 (0.3) | 7.93 (0.08) | 9.8 (0.3) | 5.9 (0.3) | 5.7 (0.3) | 1.77 (0.02) |
| 2866 | 27.9 (0.4) | 13.4 (0.1) | 10.9 (0.2) | 6.3 (0.1) | 8.4 (0.3) | 4.4 (0.1) | 4.1 (0.2) | 1.34 (0.03) |
| 4299 | 25.5 (0.2) | 12.3 (0.1) | 9.6 (0.1) | 5.73 (0.04) | 7.4 (0.1) | 3.86 (0.07) | 3.7 (0.1) | 1.13 (0.01) |
| 5732 | 24.0 | 11.5 | 8.8 | 5.25 | 6.9 | 3.34 | 3.2 | 1.00 |

| Training | **Dipole moment MAE (SD), kcal/mol** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| set size | CM | $H_{\text{core}}$ | GWH | Hückel | SAP | SAD | LB | SLATM |
| 716 | 0.357 (0.004) | 0.343 (0.005) | 0.338 (0.003) | 0.304 (0.008) | 0.272 (0.005) | 0.297 (0.004) | 0.283 (0.002) | 0.272 (0.008) |
| 1433 | 0.330 (0.002) | 0.320 (0.003) | 0.309 (0.005) | 0.290 (0.005) | 0.250 (0.007) | 0.267 (0.004) | 0.260 (0.005) | 0.218 (0.007) |
| 2866 | 0.307 (0.003) | 0.303 (0.002) | 0.290 (0.003) | 0.270 (0.003) | 0.232 (0.003) | 0.245 (0.005) | 0.233 (0.002) | 0.177 (0.005) |
| 4299 | 0.290 (0.003) | 0.295 (0.002) | 0.277 (0.001) | 0.266 (0.003) | 0.222 (0.002) | 0.231 (0.002) | 0.220 (0.004) | 0.157 (0.001) |
| 5732 | 0.281 | 0.286 | 0.270 | 0.258 | 0.217 | 0.221 | 0.211 | 0.146 |

| Training | **HOMO energy MAE (SD), eV** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| set size | CM | $H_{\text{core}}$ | GWH | Hückel | SAP | SAD | LB | SLATM |
| 716 | 0.49 (0.01) | 0.47 (0.02) | 0.41 (0.02) | 0.36 (0.01) | 0.295 (0.004) | 0.24 (0.02) | 0.23 (0.01) | 0.220 (0.009) |
| 1433 | 0.427 (0.009) | 0.397 (0.008) | 0.35 (0.01) | 0.300 (0.007) | 0.273 (0.005) | 0.194 (0.006) | 0.186 (0.006) | 0.175 (0.003) |
| 2866 | 0.383 (0.009) | 0.362 (0.005) | 0.310 (0.004) | 0.257 (0.009) | 0.238 (0.004) | 0.155 (0.002) | 0.157 (0.005) | 0.135 (0.002) |
| 4299 | 0.357 (0.003) | 0.345 (0.003) | 0.288 (0.003) | 0.230 (0.003) | 0.223 (0.002) | 0.137 (0.001) | 0.140 (0.002) | 0.114 (0.003) |
| 5732 | 0.344 | 0.331 | 0.274 | 0.217 | 0.210 | 0.127 | 0.130 | 0.099 |

| Training | **HOMO–LUMO gap MAE (SD), eV** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| set size | CM | $H_{\text{core}}$ | GWH | Hückel | SAP | SAD | LB | SLATM |
| 716 | 0.67 (0.01) | 0.63 (0.02) | 0.56 (0.02) | 0.54 (0.01) | 0.478 (0.005) | 0.44 (0.01) | 0.44 (0.01) | 0.288 (0.013) |
| 1433 | 0.60 (0.01) | 0.575 (0.007) | 0.483 (0.006) | 0.458 (0.005) | 0.43 (0.01) | 0.38 (0.01) | 0.38 (0.01) | 0.231 (0.006) |
| 2866 | 0.55 (0.01) | 0.520 (0.008) | 0.418 (0.006) | 0.393 (0.006) | 0.377 (0.004) | 0.316 (0.003) | 0.317 (0.005) | 0.182 (0.002) |
| 4299 | 0.51 (0.01) | 0.491 (0.004) | 0.391 (0.003) | 0.364 (0.005) | 0.348 (0.005) | 0.290 (0.002 | 0.294 (0.004) | 0.163 (0.001) |
| 5732 | 0.49 | 0.474 | 0.368 | 0.344 | 0.324 | 0.267 | 0.277 | 0.150 |

TABLE S2. Numerical values for the learning curves (Figure 2)

| Training set size | **Atomization energy MAE (SD), kcal/mol** | | | | | | |
|---|---|---|---|---|---|---|---|
| | LB | LB+0.001 rnd | LB+0.005 rnd | LB+0.01 rnd | LB+0.1 rnd | rnd | PBE0 |
| 716 | 8.2 (0.3) | 7.6 (0.5) | 9.4 (0.9) | 11 (2) | 42 (1) | 158 (2) | 13.2 (0.6) |
| 1433 | 5.6 (0.4) | 5.5 (0.2) | 6.6 (0.3) | 8.7 (0.2) | 36.2 (0.6) | 146.1 (0.8) | 10.3 (0.6) |
| 2866 | 4.2 (0.1) | 4.3 (0.1) | 5.3 (0.2) | 7.3 (0.2) | 30.9 (0.3) | 141.4 (1.0) | 7.8 (0.4) |
| 4299 | 3.6 (0.1) | 3.71 (0.07) | 4.8 (0.1) | 7.1 (0.2) | 28.1 (0.4) | 138.7 (0.6) | 6.6 (0.1) |
| 5732 | 3.2 | 3.30 | 4.5 | 6.6 | 25.9 | 138.3 | 5.9 |

TABLE S3. Numerical values for the learning curves (Figure 3)

| Training set size | **Atomization energy MAE (SD), kcal/mol** | | |
|---|---|---|---|
| | core | valence | full |
| 716 | 94 (1) | 18.9 (0.3) | 8.2 (0.4) |
| 1433 | 86 (1) | 15.2 (0.3) | 5.7 (0.3) |
| 2866 | 81 (1) | 12.3 (0.2) | 4.1 (0.2) |
| 4299 | 77.3 (0.5) | 10.6 (0.1) | 3.7 (0.1) |
| 5732 | 75.3 | 9.7 | 3.2 |

| Training set size | **Dipole moment MAE (SD), kcal/mol** | | |
|---|---|---|---|
| | core | valence | full |
| 716 | 0.323 (0.009) | 0.296 (0.004) | 0.283 (0.002) |
| 1433 | 0.297 (0.006) | 0.275 (0.006) | 0.260 (0.005) |
| 2866 | 0.279 (0.002) | 0.255 (0.004) | 0.233 (0.002) |
| 4299 | 0.272 (0.001) | 0.240 (0.004) | 0.220 (0.004) |
| 5732 | 0.265 | 0.229 | 0.211 |

| Training set size | **HOMO energy MAE (SD), eV** | | |
|---|---|---|---|
| | core | valence | full |
| 716 | 0.409 (0.007) | 0.218 (0.005) | 0.23 (0.01) |
| 1433 | 0.363 (0.005) | 0.184 (0.006) | 0.186 (0.006) |
| 2866 | 0.339 (0.003) | 0.158 (0.003) | 0.157 (0.005) |
| 4299 | 0.325 (0.003) | 0.144 (0.001) | 0.140 (0.002) |
| 5732 | 0.314 | 0.134 | 0.130 |

| Training set size | **HOMO–LUMO gap MAE, eV** | | |
|---|---|---|---|
| | core | valence | full |
| 716 | 0.60 (0.01) | 0.49 (0.02) | 0.44 (0.01) |
| 1433 | 0.55 (0.01) | 0.43 (0.01) | 0.384 (0.010) |
| 2866 | 0.513 (0.008) | 0.363 (0.004) | 0.317 (0.005) |
| 4299 | 0.497 (0.003) | 0.329 (0.004) | 0.294 (0.004) |
| 5732 | 0.484 | 0.306 | 0.277 |

TABLE S4. Numerical values for the learning curves (Figure 4): HOMO energy MAE (SD), eV

| Training set size | **M and $\mathbf{M}^{++}$** | | | |
|---|---|---|---|---|
| | CM | SPA$^{\mathrm{H}}$M-LB | SPA$^{\mathrm{H}}$M-LBm | SLATM |
| 720 | 6.38 (0.03) | 2.6 (0.1) | 0.40 (0.01) | 6.22 (0.03) |
| 1440 | 6.48 (0.02) | 1.7 (0.1) | 0.337 (0.004) | 6.29 (0.02) |
| 2880 | 6.66 (0.01) | 1.13 (0.06) | 0.286 (0.003) | 6.30 (0.03) |
| 4320 | 6.73 (0.02) | 0.87 (0.03) | 0.260 (0.004) | 6.320 (0.006) |
| 5760 | 6.85 | 0.70 | 0.244 | 6.334 |

| Training set size | **M and $\mathbf{M}^{+\cdot}$** | | | |
|---|---|---|---|---|
| | CM | SPA$^{\mathrm{H}}$M-LB | SPA$^{\mathrm{H}}$M-LBm | SLATM |
| 720 | 3.35 (0.01) | 1.3 (0.1) | 0.35 (0.01) | 3.27 (0.02) |
| 1440 | 3.38 (0.01) | 0.86 (0.06) | 0.31 (0.01) | 3.27 (0.02) |
| 2880 | 3.46 (0.01) | 0.60 (0.02) | 0.273 (0.004) | 3.29 (0.01) |
| 4320 | 3.515 (0.008) | 0.498 (0.005) | 0.257 (0.004) | 3.298 (0.004) |
| 5760 | 3.569 | 0.457 | 0.242 | 3.302 |

| Training set size | **M, $\mathbf{M}^{+\cdot}$, and $\mathbf{M}^{++}$** | | | |
|---|---|---|---|---|
| | CM | SPA$^{\mathrm{H}}$M-LB | SPA$^{\mathrm{H}}$M-LBm | SLATM |
| 1080 | 4.76 (0.02) | 1.87 (0.09) | 0.400 (0.008) | 4.52 (0.01) |
| 2160 | 4.71 (0.03) | 1.21 (0.04) | 0.340 (0.006) | 4.47 (0.02) |
| 4320 | 4.71 (0.01) | 0.76 (0.03) | 0.295 (0.003) | 4.40 (0.02) |
| 6480 | 4.70 (0.01) | 0.53 (0.02) | 0.276 (0.002) | 4.38 (0.01) |
| 8640 | 4.68 | 0.41 | 0.260 | 4.36 |

TABLE S5. Numerical values for the learning curves (Figure 5): dipole moment MAE (SD), a.u.

| Training set size | SLATM | SPA$^{\mathrm{H}}$M-LB(m) |
|---|---|---|
| 557 | 0.94 (0.06) | 0.71 (0.02) |
| 1111 | 0.79 (0.02) | 0.65 (0.01) |
| 2223 | 0.68 (0.01) | 0.588 (0.008) |
| 4334 | 0.62 (0.01) | 0.561 (0.008) |
| 5446 | 0.600 | 0.520 |

TABLE S6. Numerical values for Figure 6

| | **QM7** | | **L11** | |
|---|---|---|---|---|
| | SLATM | SPA$^{\mathrm{H}}$M-LB(m) | SLATM | SPA$^{\mathrm{H}}$M-LB(m) |
| time to compute the representation, s | $5.38 \cdot 10^1$ | $4.79 \cdot 10^2$ | $1.55 \cdot 10^3$ | $8.01 \cdot 10^1$ |
| time to compute the kernel, s | $5.38 \cdot 10^1$ | $2.07 \cdot 10^0$ | $2.93 \cdot 10^2$ | $1.37 \cdot 10^0$ |

## REFERENCES

[S1]G. Knizia, J. Chem. Theory Comput. **9**, 4834 (2013).
[S2]L. Amat and R. Carbó-Dorca, Int. J. Quantum Chem. **87**, 59 (2001).
[S3]J. H. Van Lenthe, R. Zwaans, H. J. J. Van Dam, and M. F. Guest, J. Comput. Chem. **27**, 926 (2006).
[S4]M. Cordova, M. D. Wodrich, B. Meyer, B. Sawatlon, and C. Corminboeuf, ACS Catal. **10**, 7021 (2020).
[S5]J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio, and A. Tkatchenko, Sci. Data **8**, 43 (2021).