

Exploring the Design Space of Machine Learning Models for Quantum Chemistry with a Fully Differentiable Framework

Divya Suman,[#] Jigyasa Nigam,^{*,#} Sandra Saade, Paolo Pegolo, Hanna Türk, Xing Zhang, Garnet Kin-Lic Chan, and Michele Ceriotti^{*}



Cite This: *J. Chem. Theory Comput.* 2025, 21, 6505–6516



Read Online

ACCESS |

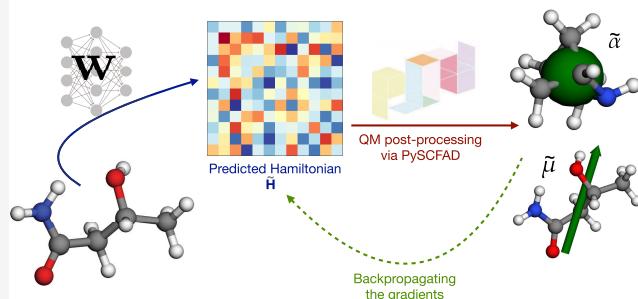
Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Traditional atomistic machine learning (ML) models serve as surrogates for quantum mechanical (QM) properties, predicting quantities such as dipole moments and polarizabilities directly from compositions and geometries of atomic configurations. With the emergence of ML approaches to predict the “ingredients” of a QM calculation, such as the ground-state charge density or the effective single-particle Hamiltonian, it has become possible to obtain multiple properties through analytical physics-based operations on these intermediate ML predictions. We present a framework that seamlessly integrates the prediction of an effective electronic Hamiltonian, for both molecular and condensed-phase systems, with PySCFAD, a differentiable QM workflow. This integration facilitates training models indirectly against functions of the Hamiltonian, such as electronic energy levels, dipole moments, polarizability, etc. We then use this framework to explore various possible choices within the design space of hybrid ML/QM models, examining the influence of incorporating multiple targets on model performance and learning a reduced-basis ML Hamiltonian that can reproduce targets computed on a much larger basis. Our benchmarks evaluate the accuracy and transferability of these hybrid models, compare them against predictions of atomic properties from their surrogate models, and provide indications to guide the design of the interface between the ML and QM components of the model.

Differentiable and modular framework for Hamiltonian learning



Downloaded via NATL UNIV OF SINGAPORE on August 16, 2025 at 04:43:10 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

1. INTRODUCTION

Machine learning (ML) has become indispensable for atomistic modeling of molecules and materials, driving scientific discovery and accelerating the search for compounds with desired properties. ML approaches have not only enabled large-scale molecular dynamics through accurate predictions of potential energy surfaces^{1–9} but also refined our understanding of the intricate relationships between atomic geometries and physical observables, including electronic properties such as dipole moments^{10,11} and polarizabilities.^{12–15}

Although these surrogate models can describe complex molecular behaviors, they often cannot infer properties beyond those they were trained to reproduce or struggle to extrapolate to structures that deviate considerably from those included in the training set.^{10,12} One possible approach to address these limitations is to develop models that aim for broader transferability by learning fundamental quantities central to the electronic structure problem, such as the electron density^{16–21} or N -electron density matrices,^{22–24} the electronic wave function^{25–30} or representations of an effective single-particle Hamiltonian operator in a specified atomic orbital (AO) basis.^{31–34} Learning any of these underlying electronic quantities at a given level of theory grants access, in principle,

to all observables that can be obtained from relatively inexpensive postprocessing operations. In this work, we focus on the problem of learning an effective single-particle Hamiltonian.

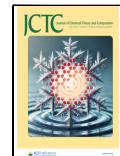
The matrix elements of Hamiltonian \mathbf{H} encode pairwise interactions between the AO basis functions, which can be centered on either the same atom (constituting on-site interactions) or two different atoms (off-site interactions) in a given molecular configuration. A key property of \mathbf{H} is its equivariance under operations of the $O(3)$ group of rotations and inversions, as well as under permutations of identical atoms.³⁵ Early ML approaches for modeling \mathbf{H} often circumvented explicitly describing these symmetries by using data augmentation,³¹ relying on ad-hoc modifications of atom-centered descriptors,³⁶ or by targeting observables such as optical excitations³⁷ through an intermediate *invariant*

Received: April 1, 2025

Revised: June 6, 2025

Accepted: June 16, 2025

Published: June 25, 2025



representation of the Hamiltonian with respect to these geometric transformations. More recent approaches handle symmetries directly and account for pair dependence of matrix elements using bond-centered atomic cluster expansion descriptors³⁴ or through equivariant features describing the corresponding atom pairs.³⁵ Equivariant approaches based on message-passing neural networks^{33,38–42} have similarly risen to the task, as their underlying equivariant node and edge features can be naturally adapted to learn pairwise quantities.

Within a framework for Hamiltonian machine learning, there are several possibilities to define the target. For instance, one may learn the exact matrix representation corresponding to a mean-field calculation performed by using a specified electronic structure method and basis set, accepting the fact that any truncated basis will introduce a finite-basis-set error. An alternative would be to learn a reduced effective \mathbf{H} that reproduces observables from a higher level of theory or a larger-basis calculation. Using this strategy of integrating ML and quantum chemistry, ref 43 optimizes an effective minimal-basis \mathbf{H} to match properties (eigenspectrum, Löwdin charges) derived from quantum mechanical (QM) calculations performed on a much larger basis set. A similar framework⁴⁴ refines an effective one-electron Hamiltonian by machine learning a correction so that multiple properties derived from it align with reference data computed using higher-accuracy many-body perturbative methods, such as CCSD. Such hybrid approaches have demonstrated improved accuracy and transferability of ML at a lower computational cost. Interfacing ML approaches with quantum chemistry is made simpler by the incorporation of automatic differentiation (AD) capabilities within electronic structure codes.⁴⁵ While this effort is useful in its own right, as it allows one to compute properties that were previously inaccessible due to the absence or impracticality of analytical derivatives, it is especially useful in combination with ML frameworks designed around evaluating gradients of a loss function through backpropagation and has been explored in several recent works.^{46–48} By training ML models on physical quantities derived from electronic structure calculations in a fully differentiable framework, one can attempt to improve predictive accuracy and enhance transferability across chemical systems.

Developing effective ML models for intermediate descriptions of the electronic structure requires accounting for the subtleties of quantum chemistry (including the choice of the level of theory and basis set) in addition to conventional ML tasks (such as selecting appropriate architectures). In this work, we combine our previously proposed hybrid (or indirect) ML architecture⁴³ with the autodifferentiable electronic structure code PySCFAD⁴⁹ to optimize an intermediate representation of \mathbf{H} that reproduces target electronic properties derived through explicit QM manipulation of a machine-learned effective Hamiltonian. In particular, we target molecular dipole moments (μ) and polarizabilities (α), as well as band energies in condensed-phase systems. A key challenge in training models on downstream QM observables is that these quantities are typically nonlinear functions of the Hamiltonian, involving matrix diagonalizations and calculations of responses to perturbations. Optimizing model parameters to minimize the loss on these observables, therefore, requires underlying operations to be implemented in a differentiable manner. PySCFAD, a differentiable electronic structure engine, readily provides this capability, eliminating the need to implement

custom routines for each property of interest. By embedding PySCFAD within the training pipeline, our ML framework enables the intermediate ML Hamiltonian to be tuned such that it accurately reproduces the target QM observables through appropriate operations, allowing for end-to-end optimization.

For our study, we keep a minimalistic symmetry-adapted parametrization of \mathbf{H} and systematically explore how variations in quantum chemical design choices affect the performance of the model by progressively increasing the physical constraints in our model, retaining a diverse subset of the QM7 data set^{50–52} for training and evaluating its generalizability to larger systems, including the QM9^{53,54} data set. Our results demonstrate that a well-constrained indirect model often improves predictive accuracy for both QM7 and QM9 data sets, especially for response properties such as polarizability. When trained to reproduce properties from larger-basis set calculations, the effective indirect minimal-basis models achieve an accuracy comparable to that of their large-basis counterparts while remaining more computationally efficient. Even if the indirect minimal-basis models are less accurate when trained against properties from large-basis set calculations than when trained against results from a minimal basis calculation their accuracy is much better than the basis set error itself. We show that it is, therefore, preferable to train against large-basis outputs, as the predictions are more accurate in an absolute sense, even though the model has the size and computational cost of a minimal basis model. We also briefly discuss the influence of the effective Hamiltonian basis size.

We further evaluate the model's transferability on larger, more complex molecules than the training set, where the property-specific surrogate models have previously been limited due to their inability to capture nonlocal effects.¹⁰ Although this work focuses on a specific model architecture and training strategies, the overall ML framework is modular and extensible, allowing for flexible substitution of atomic descriptors, model architectures, and differentiable electronic structure routines. For example, the input descriptors can be replaced by higher-body-order features^{55,56} or other widely used atomic representations,^{8,34} and the complexity of the model can be adjusted by varying the number and width of layers or incorporating other nonlinearities. Alternatively, this component can be replaced entirely by graph neural networks that operate directly on atomic coordinates. Likewise, the differentiable QM backend can be swapped out for other codes that offer functionality similar as PySCFAD. This modular design contributes to the broader effort of integrating ML with QM, with models that combine the interpretability and transferability of physics-based approaches with the efficiency of data-driven methods.

2. THEORY AND METHODS

In many quantum chemistry frameworks, the effective single-electron wave function is expanded in terms of localized orbital basis functions centered on atomic positions (atomic orbitals, AO). These basis functions are typically nonorthogonal due to their localized nature. The single-electron eigenstates are obtained by solving a generalized eigenvalue equation including the overlap matrix between the basis functions, S ,

$$\mathbf{H}\mathbf{C} = \mathbf{S}\mathbf{C} \operatorname{diag} \boldsymbol{\epsilon}, \quad (1)$$

where \mathbf{e} is the vector of one-electron eigenvalues, or molecular orbital (MO) energies, and \mathbf{C} is the (unitary) matrix of MO coefficients. In mean-field theories, \mathbf{H} depends self-consistently on the MOs, but in the following, we will assume that \mathbf{C} can be obtained from a single diagonalization. The one-electron density matrix ρ is then computed from the MO coefficients and the occupation numbers \mathbf{f} of the molecular orbitals as

$$\rho = \mathbf{C} \operatorname{diag}(\mathbf{f}) \mathbf{C}^\dagger. \quad (2)$$

Any ground-state property can then be computed from the MO energies and the density matrix. For instance, the total dipole moment μ of an atomic structure A can be computed as

$$\mu(A) = -e \operatorname{Tr}(\rho_A \mathbf{x}_A) + \sum_{i \in A} Q_i \mathbf{r}_i. \quad (3)$$

The first term is the electronic contribution to the dipole moment, where e is the electronic charge and \mathbf{x}_A is the representation of the electronic position operator $\hat{\mathbf{x}}$ in the AO basis of structure A , while the second term is the nuclear contribution, computed as the product of the effective nuclear charges, Q_i , and positions, \mathbf{r}_i , of each atom i . Response properties, in turn, are obtained in the linear approximation as the derivatives of ground-state observables. The polarizability tensor α , for instance, can be computed as the (zero-field) derivative of the dipole with respect to an applied electric field. Thus, modeling the Hamiltonian matrix \mathbf{H} enables the prediction of a wide range of electronic properties.

It is possible to train ML models to directly predict each property starting from atomic species and coordinates without explicitly modeling the Hamiltonian. This scheme, however, requires a separate ML model to be trained for each property of interest. We use these *direct property models* as a baseline for comparison with effective Hamiltonian models that provide a physically interpretable intermediate from which various properties can be derived simultaneously. As we shall see, this second approach also offers better transferability to out-of-sample molecules. As explained below, we use simple linear models to parametrize \mathbf{H} , even though they are not especially accurate, in order to focus our attention on the role played by the many choices one can make in the design of the part of the model that more closely follows the structure of explicit QM calculations.

2.1. ML Models of Atomistic Properties. The dipole moment is usually represented as a Cartesian tensor of rank one, whereas polarizability is a Cartesian tensor of rank two. Although it is possible to model these properties in this form, it is usually more convenient to decompose them into irreducible representations (irreps) of $O(3)$, each of which can be individually modeled, as in refs 10,12,

$$y_A \equiv \bigoplus_{\sigma, \lambda} y_A^{\sigma \lambda}. \quad (4)$$

In other words, the Cartesian representation of the property y for structure A , y_A , is expressed as a direct sum of $O(3)$ irreps indexed by $\sigma \lambda$, each with a behavior under spatial inversion denoted by σ , and rotational behavior as a spherical harmonic $Y_\lambda^\mu(\hat{\mathbf{x}})$, which can be enumerated as a vector of size $2\lambda + 1$, with $\mu \in [-\lambda, \lambda]$. For example, the dipole moment being a vector decomposes into a single irrep indexed by $\lambda = 1, \sigma = 1$, whereas the polarizability, being a symmetric tensor, decomposes into irreps labeled by (λ, σ) pairs corresponding to $(0, 1), (2, 1)$. Note that we use $\sigma = 1$ to denote the object has a parity similar

to a polar tensor corresponding to the same λ , whereas $\sigma = -1$ denotes a pseudotensor.³⁵

These properties are modeled as additive quantities, that is, as sums of atomic contributions and can be conveniently approximated in terms of atom-centered descriptors that exactly mirror the (improper) rotational nature of the target, combined with invariant weights,

$$y_A^{\sigma \lambda \mu} = \sum_{i \in A} y_{A_i}^{\sigma \lambda \mu} \approx \sum_{i \in A} \mathbf{w}^{\sigma \lambda a_i} \cdot \xi_{A_i}^{\sigma \lambda \mu}. \quad (5)$$

$\xi_{A_i}^{\sigma \lambda \mu}$ denotes the equivariant λ -SOAP³⁷ descriptor for atom i in structure A , where a_i is the atomic species of i . The invariant model weights $\mathbf{w}^{\sigma \lambda a}$ (invariance implies that the same regression weights apply to all μ components) are indexed by λ, σ , and a to highlight that each target irrep can be learned by a distinct linear model. The dot-product in eq 5 is taken over the feature dimension, which includes radial and angular components of the basis used to compute the λ -SOAP descriptor, and information about the chemical variability in the environment centered on i . In the following, we will use these kinds of symmetry-adapted regression models as examples of property models for dipoles and polarizabilities, analogous to the kernel models used in MuML¹⁰ and AlphaML.¹² The choice of a simple ridge regression model is made in the same spirit as the restriction of the Hamiltonian model to a linear form, keeping a minimalistic form of ML with a restricted design space to focus on the effect of the “QM-facing” part of the model architecture.

2.2. ML Models of Effective Single-Particle Hamiltonians. In contrast to global properties such as dipoles and polarizabilities, which are modeled as a sum of atom-centered contributions but are trained against references computed for the entire molecular structure, the Hamiltonian matrix elements depend on specific pairs of orbitals involved in the interaction. When these orbitals are centered on atoms, as is the case for localized AO bases, the Hamiltonian matrix elements can be viewed as objects labeled by pairs of atoms, as well as multiple quantum numbers, namely, the radial (n) and the angular (l, m) quantum numbers characterizing each AO. These angular functions are typically chosen to be (real) spherical harmonics and determine the equivariant behavior of the matrix elements under rotations and inversions. The nonequivalence of matrix elements under the exchange of atom labels, while keeping the orbitals fixed, also makes them equivariant under permutations of the atom labels.³⁵

In the same spirit of equivariant models of global properties, we describe the rotational behavior of \mathbf{H} by transforming each pair of angular functions into symmetrized irreps of $O(3)$. For each pair of angular quantum numbers (l, l') associated with the radial labels n and n' for atoms i and j , we couple the angular functions using Clebsch–Gordan coefficients to obtain equivariant outputs indexed by $\lambda \in [l - l', l + l']$,

$$\mathbf{H}_{A_{ij}}^{\mathbf{p} \sigma \lambda \mu} = \sum_{mm'} \langle lm; l'm' | \lambda \mu \rangle \mathbf{H}_{A_{ij}}^{\mathbf{p} mm'}, \quad (6)$$

where we use the shorthand $\mathbf{p} = (n, l, n', l')$ to denote the combined set of indices for the angular and radial basis functions, as well as the chemical species of the atoms, and $\langle lm; l'm' | \lambda \mu \rangle$ are the Clebsch–Gordan coefficients.

To address the two-centered nature of the matrix elements and construct a model akin to eq 5, we extended atom-centered descriptors in ref 35. to a framework capable of

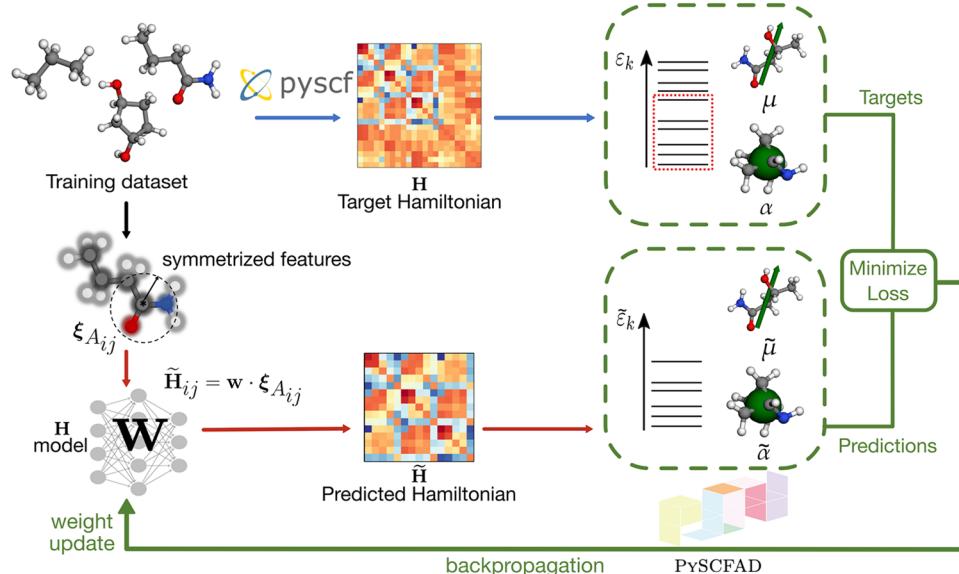


Figure 1. Schematic workflow depicting the modular integration of the ML prediction of electronic Hamiltonians (for a selected basis set) with PySCFAD. The upper branch of the workflow demonstrates the generation of the reference Hamiltonian corresponding to an atomic configuration through PySCF, from which observables, including the dipole moment (μ), polarizability (α), and MO energies (ϵ_k), are computed. These quantities serve as the targets for various models described in the text. Symmetry-adapted pair features $\xi_{A_{ij}}$ computed for each pair of atoms (i, j) in each molecular configuration (A) are used as inputs to an ML model that yields a prediction of the Hamiltonian matrix elements (\tilde{H}_{ij}). When interfaced with PySCFAD, these predictions can be used to compute the corresponding predictions of molecular properties ($\tilde{\mu}$, $\tilde{\alpha}$, and $\tilde{\epsilon}_k$). The model can be optimized to minimize the loss directly on the Hamiltonian prediction or, alternatively, minimize the loss on the molecular properties derived from it. The red and green arrows correspond to the forward and backward passes of the model, respectively.

describing multiple atomic centers and their connectivities, giving rise to the equivariant pair descriptor $\xi_{A_{ij}}^{\sigma\lambda\mu}$, which simultaneously characterizes the environments of atoms i and j in structure A . The features are built as a symmetrized product of density expansion coefficients for atomic pairs and atom-centered neighbor densities and were also used in ref 43 (See the Supporting Information (SI)) for a full discussion of the features and their hyperparameters. Here, $\lambda\mu$ denotes the rotational symmetry, and σ indicates inversion parity, as in eq 5. Each Hamiltonian block (eq 6) can be modeled separately, for instance, through a linear layer,

$$\mathbf{H}_{A_{ij}}^{\sigma\lambda\mu} = \mathbf{w}^{\sigma\lambda} \xi_{A_{ij}}^{\sigma\lambda\mu} + \delta_{\lambda 0} b^{\sigma}, \quad (7)$$

where $\mathbf{w}^{\sigma\lambda}$ is an invariant weight and the intercept $\delta_{\lambda 0} b^{\sigma}$ is nonzero only for invariant blocks, to maintain equivariance.

2.3. Electronic Hamiltonians as Elements in an ML Architecture. Matrix elements of the effective one-electron Hamiltonian have not only served as targets for machine learning but also as inputs to models that have successfully predicted a wide range of molecular properties.^{58–60} Hybrid ML-QM models treat \mathbf{H} in yet another way, using an ML prediction of the Hamiltonian as an intermediate component in a modular pipeline that computes the desired chemical properties, given an input molecular configuration.⁴³ Rather than explicitly targeting the matrix elements obtained from a QM calculation, the Hamiltonian is learned implicitly as an intermediate layer used to predict several physical properties accessible in a quantum chemistry calculation, which are used as the targets for ML. The implicit Hamiltonian does not need to be expressed on the same basis as used for calculating reference values of these target properties, enabling the emulation of large basis-set calculations with a significantly simpler and smaller model. We denote this learnt representa-

tion of the Hamiltonian on an effective basis (EB) as \mathbf{H}_{EB} . In ref 43, this procedure demonstrated promising results in both extrapolating to diverse structures, including molecules much larger than the training set, and generalizing to observables not optimized during training, such as molecular excitations.

To extend the capabilities of these models beyond direct diagonalization and facilitate the computation of observables that require nontrivial manipulations of the Hamiltonian, we interface our ML model with PySCFAD, an electronic structure code that supports automatic differentiation,⁴⁹ as shown schematically in Figure 1. The Hamiltonian matrix predicted from eq 7 is passed to PySCFAD, which computes QM observables by diagonalizing the matrix and evaluating the corresponding expectation values, thus delegating the physics-based calculations to dedicated electronic structure software. PySCFAD's automatic differentiation capabilities enable the optimization of losses on these target observables with respect to the model parameters by propagating the gradients through the predicted Hamiltonian. The availability of such a modular framework enables the investigation of the accuracy and transferability of these hybrid models and how they depend on the choice of learning targets and, more generally, on the details of the part of the model architecture that is explicitly built to mimic physics-based manipulations of \mathbf{H} .

3. RESULTS

3.1. Data Set Overview and Model Classes. For training, we use a subset of 700 molecules from the QM7b data set,^{51,52} containing only molecules composed of C, H, N, and O atoms. These molecules were selected from a total of 9000 using farthest point sampling⁶¹ on three-body atom-centered descriptors (SOAP⁶²) to maximize structural diversity. Details of sample selection, reference electronic

structure calculations, and generation of the data set are provided in [Section S4](#) in the SI. We use calculations performed with three different basis sets: a minimal later type orbital (STO)-3G basis, a larger def2-TZVP basis, as well as an intermediate-size 6–31G basis that we use to test the impact of the effective basis set size. The accuracy of the model is evaluated on two separate test sets, the first consisting of 100 QM7 molecules, and the second composed of 200 QM9 molecules,^{53,54} both spanning the same compositional space. In addition, we assess performance on a few targeted benchmarks, as discussed below.

We compare four distinct modeling strategies, as summarized in [Figure 2](#), each differing in how they embed physical

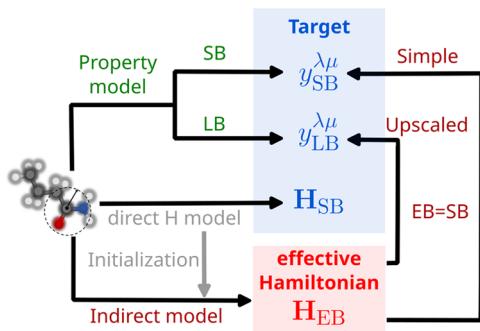


Figure 2. Overview of various modeling strategies to combine QM and ML. Traditional atomistic ML approaches learn directly the map between input atom-centered descriptors and molecular properties, such as dipole moments and polarizabilities (property model). Here, we investigate indirect models that optimize an effective Hamiltonian (H_{EB}) to simultaneously recover multiple observables (indicated as $y^{\lambda\mu}$) via differentiable operations in PySCFAD, either in an effective basis (simple indirect) or in a larger basis LB (upscaled indirect). In the spirit of transfer learning, both models are initialized from a direct Hamiltonian model (H_{SB}), pretrained to reproduce minimal-basis Hamiltonian matrix elements using linear regression on the same descriptors. In the examples presented in [Section 3](#), we use the minimal basis STO-3G as the SB and EB.

(QM) knowledge within ML and define learning targets. To reiterate, property models predict molecular properties such as dipole moments or polarizabilities directly from atomic configurations using atom-centered descriptors as defined in [eq 5](#). Although widely used and well-established, these models require separate training for each property, which limits their transferability. The second approach is a direct Hamiltonian model, which we refer to as the direct \mathbf{H} model. This model is also trained using equivariant generalizations of atom-centered descriptors to reproduce the matrix elements of a minimal basis (or a small basis) single-particle Hamiltonian (H_{SB}), [eq 6](#). We can compute the electronic properties from the predicted Hamiltonians although these properties are not optimized. The remaining two approaches constitute indirect models, where the Hamiltonian serves as an intermediate layer rather than as the target. Both indirect models parametrize an effective basis representation of the Hamiltonian (H_{EB}) but differ in training objectives. Simple indirect models restrict the optimized effective Hamiltonian to have the same basis label (EB) as the one used to compute reference electronic properties (SB). Upscaled models, in contrast, target properties computed using the larger def2-TZVP basis (LB), thus combining the computational efficiency of an effective basis parametrization with the accuracy of a larger one. In both cases, we choose the

effective basis to coincide with SB for simplicity, but we provide an example in [SI](#) where EB differs from both SB and LB. In the following sections, we report the performance of indirect models and compare them against property models, highlighting that their increased flexibility improves predictive accuracy when paired with appropriate physics-based constraints.

3.2. Design Space of Effective Hamiltonian Models.

When using a model that explicitly learns the matrix elements of a target \mathbf{H} in a given AO basis set, one does not need to separately learn the AO representation of the overlap matrix or that of operators needed to compute other properties (e.g., the position operator), as they can be computed inexpensively in the same basis. In contrast, when using an indirect model that learns one or more derived properties—and particularly, when targeting property values computed with a different basis—the intermediate Hamiltonian learned within the model does not correspond to a well-defined basis (as discussed further below). Consequently, the overlap matrix and the AO representation of various operators must also be redefined for consistency. These could either be considered as additional components within the model, having the same size and symmetries compatible with the parametrization of \mathbf{H} , but could originate from any compatible basis set or be learned separately. We experimented with learnable overlap and other operators but found that this approach led to model instabilities as additional constraints (e.g., enforcing the overlap to be positive definite) were necessary to maintain physical consistency and numerical stability.

In the following, unless stated otherwise, we use a linear model and the STO-3G basis featuring a minimal number of AOs on each atom as the model basis. To compute functions of the Hamiltonian, we use the reference STO-3G representation of the overlap, as well as the AO representation of all operators computed on this basis. To improve the model convergence, we initialize the model weights to those obtained from a symmetry-adapted ridge regression (7) model (direct \mathbf{H} model). These pretrained model weights are subsequently refined using stochastic gradient descent on specific target properties derived from predicted \mathbf{H} . More details about the training procedures are provided in [Supporting Information](#) (SI).

We assess the performance of two classes of indirect Hamiltonian models, *simple* and *upscaled* indirect models, as additional target properties are gradually introduced. For each model class, we progressively include the following target properties during optimization: MO energies (ϵ), molecular dipole moments (μ), molecular polarizabilities (α), and Mayer bond order (B).⁶³ The Mayer bond order is defined as

$$B_{ij} = \sum_{\eta \in i} \sum_{\eta' \in j} (\rho S)_{\eta\eta'} (\rho S)_{\eta'\eta}, \quad (8)$$

where η and η' label AOs centered on atoms i and j , respectively. It is representative of the localization of the density matrix in the AO basis used for the target calculations while remaining independent of the choice of the specific AO labels. In the following, these fine-tuned models are denoted by the properties on which they are optimized, for instance, (ϵ) indicates a model trained solely on MO energies, while (ϵ, μ) refers to the one trained on both MO energies and dipole moments and so on.

[Figure 3](#) illustrates the relative performance of different models on the QM7 and QM9 test data sets. The top row

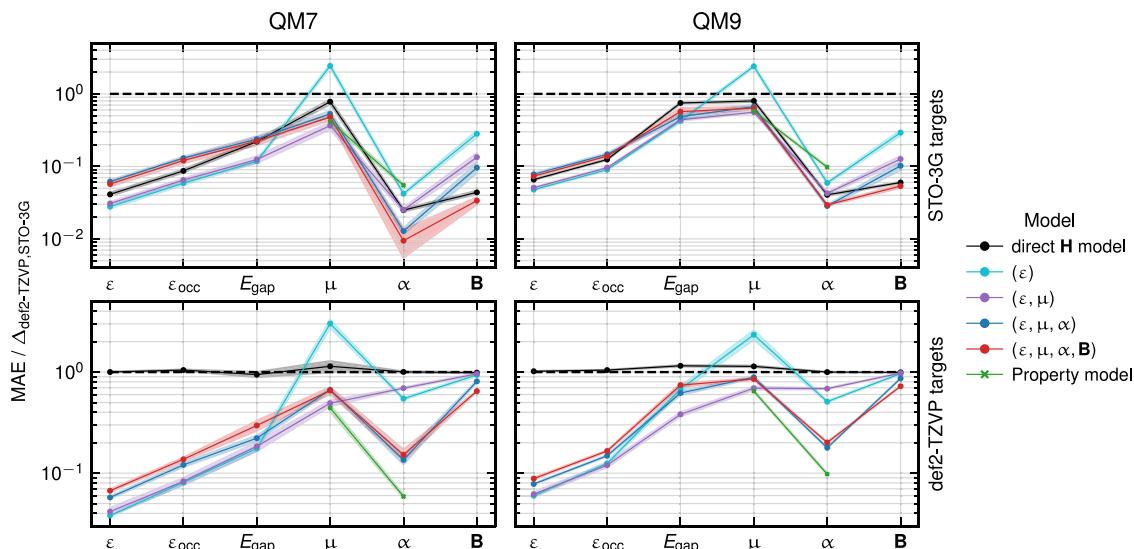


Figure 3. Comparison of model accuracies on QM7 and QM9 test data sets for various molecular properties (listed on the *x*-axis) computed using STO-3G (top) and def2-TZVP (bottom) basis sets. Each plot shows the ratio of MAEs from the prefitted ridge regression model ($\mathbf{H}_{\text{STO-3G}}$) and different indirect models (denoted by the properties they are optimized on in parentheses) to the basis set error ($\Delta_{\text{def2-TZVP,STO-3G}}$), which is the mean absolute error between the reference values computed in def2-TZVP basis and the STO-3G basis. This basis-set error serves as a baseline to compare different models and is indicated by the black, dashed line at $y = 1$, and corresponds to (for QM7 and QM9, respectively) 4.99 and 4.87 eV for all MO energy levels (ϵ), 2.86 and 3.04 eV for the occupied energy levels ϵ_{occ} , 2.60 and 1.36 eV for the HOMO–LUMO gap E_{gap} , 0.017 a.u and 0.023 a.u for the dipole moment μ , 5.66 a.u and 5.82 a.u for the polarizability α , and 5.93 a.u and 7.26 for the Mayer bond order B .

shows results for the simple indirect models, whereas the bottom row corresponds to upscaled indirect models, where the targets are computed using the def2-TZVP basis. We evaluate the performance of different models across the prediction of several properties, namely, MO energies, occupied-state MO energies (ϵ_{occ}), dipole moments, polarizabilities, highest occupied molecular orbital–lowest unoccupied molecular orbital (HOMO–LUMO) gaps (E_{gap}), and Mayer bond orders, as listed on the *x*-axis. The normalized mean absolute error (MAE) of each model is reflected on the *y*-axis. Note that these are normalized by the *basis set error* ($\Delta_{\text{def2-TZVP,STO-3G}}$), which is defined as the MAE between the reference observables computed in the def2-TZVP and the STO-3G basis, such that $y = 1$ indicates the magnitude of error associated with the convergence of the basis set. As expected, for both the QM7 and QM9 test sets, the prediction accuracy of a property is improved if it is explicitly included in the optimization. However, imposing additional constraints may impact accuracy due to the redistribution of the model flexibility and available training data over several optimization tasks. Overall, the results indicate that models optimizing all properties (ϵ , μ , α , and B) are the most robust. With the exception of the prediction of dipole moments from (ϵ) models, almost all model errors are between one and two orders of magnitude smaller than the basis set error.

Simple indirect models incur much lower MAEs as their implicit Hamiltonian representation is consistent with that of the target, and the model does not have to compensate for basis set convergence—especially for properties such as α and B that exhibit a strong basis-set sensitivity. The large basis-set error for these properties is the reason why the normalized MAE on α is always lower than that on μ , despite their absolute MAEs being comparable, as reported in Tables S3 and S4 in the SI. Similarly, the error on occupied-state MO energies is lower than that on the whole eigenspectrum because the occupied states are less sensitive to basis set

convergence than virtual state MO energies. However, upscaled indirect models need to effectively extrapolate beyond the STO-3G basis to match the def2-TZVP targets, leading to larger ML errors. This implies that basis set incompleteness is a dominant source of error and that ML struggles to fully correct for it. We also compared these models to the *property models* (5), trained on the same training data set as the indirect models to predict dipole moments and polarizability computed in the two different bases. Simple indirect models improve the prediction accuracy of polarizability on the STO-3G basis compared to the property models on the same basis. However, the property models for both dipoles and polarizabilities in the def2-TZVP basis consistently exhibit lower errors than the indirect upscaled models. Thus, even though it is possible to reproduce molecular properties computed with a large basis using the indirect upscaled model with a minimal basis representation of the Hamiltonian, these predictions may be inherently limited by the representability of the model basis. The size of the effective model basis, EB, must therefore be considered a part of the model hyperparameters. In fact, an upscaled indirect model might not be capable of representing the symmetry or overall shape of some important MOs (as illustrated through several examples in ref 43). As discussed in Section 4 of the SI, where we consider the example of 6-31G as EB, representing the effective Hamiltonian on a larger basis can improve the accuracy of an upscaled model; however, this increased expressivity may come at the cost of reduced transferability due to a higher risk of overfitting.

Another important factor affecting the accuracy of the response properties is the accuracy of the virtual MOs. Small basis sets often fail to adequately represent the virtual states due to the insufficient availability of polarization functions. According to standard linear response theory,⁶⁴ the first-order correction to the density matrix under an external electric field \mathcal{E} is given by $\Delta\rho_{\eta\nu} = \mathcal{E}\cdot\chi_{\eta\nu}$, where the response vector $\chi_{\eta\nu}$ is defined by

$$\chi_{\eta\nu} = 2 \sum_{r \in \text{occ}} \sum_{s \in \text{virt}} \frac{\mathbf{x}_{rs}}{\epsilon_r - \epsilon_s} (C_{\eta s} C_{\nu r} + C_{\eta r} C_{\nu s}). \quad (9)$$

Here, η, ν index AOs, r index MOs, \mathbf{x}_{rs} denotes the matrix elements of the electronic position operator, and \mathbf{C} is the MO coefficient matrix. As the denominator depends on virtual MO energies, enhancing their accuracy directly improves the prediction of polarizability and other linear response properties.

3.3. Interpreting Effective Hamiltonians. Even when an indirect model is represented in the same basis as its reference calculations (i.e., simple indirect models), the predicted $\tilde{\mathbf{H}}$ at the end of the training does not need to exactly match the target \mathbf{H} . The matrices can differ because of a mismatch in the eigenvalues or because of a different alignment of the eigenvectors. In order to differentiate between the two effects, we define an eigenbasis alignment (EA) operation,

$$\tilde{\mathbf{H}}_{\text{EA}} = \mathbf{C} \tilde{\mathbf{C}}^\dagger \tilde{\mathbf{H}} \tilde{\mathbf{C}} \mathbf{C}^\dagger, \quad (10)$$

$\tilde{\mathbf{C}}$ and \mathbf{C} are the eigenvector matrices of $\tilde{\mathbf{H}}$ and \mathbf{H} , respectively (we consider for simplicity Löwdin-orthogonalized matrices). If the predicted eigenspectrum exactly matches the reference one, $\tilde{\mathbf{H}}$ and \mathbf{H} are related by an orthogonal transformation, and $\tilde{\mathbf{H}}_{\text{EA}} = \mathbf{H}$. In the general case, where predictions have some error, eigenbasis alignment can still be considered the best orthogonal transformation relating $\tilde{\mathbf{H}}$ and \mathbf{H} , in the sense that the discrepancy between $\|\mathbf{H} - \tilde{\mathbf{H}}\|$ and $\|\mathbf{H} - \tilde{\mathbf{H}}_{\text{EA}}\|$ as a measure of how much indirect training rotates the representation of the predicted Hamiltonians with respect to the original basis set.

For upscaled models, where the target properties are computed using LB, which is larger than the EB used in the model, eigenbasis-aligned predictions can be compared with symmetry-adapted projected Hamiltonians (SAPHs)³⁵ defined as,

$$\mathbf{H}_{\text{SAPH}} = \mathbf{C} \overline{\mathbf{C}}_{\text{LB}}^\dagger \mathbf{H}_{\text{LB}} \overline{\mathbf{C}}_{\text{LB}} \mathbf{C}^\dagger, \quad (11)$$

where $\overline{\mathbf{C}}_{\text{LB}}$ is the submatrix of eigenvectors of \mathbf{H}_{LB} corresponding to the number of molecular orbitals described by EB, that is, the STO-3G basis, and \mathbf{C} is the same as in eq 10. We compare these quantities in Figure 4, which illustrates the deviations of $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}}_{\text{EA}}$ from the reference Hamiltonian \mathbf{H} . As expected, the deviations after EA (shown by the red curve) are significantly smaller than those of the original ML outputs (shown in blue). For simple indirect models (top panel), increasing the number of constraints directly correlates with decreased deviations from the reference in the case of the original ML outputs, revealing some correlation between a larger number of constraints and similarity between $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}}_{\text{EA}}$. Given the large number of target observables required to fully constrain the flexibility of representation and the fact that we only have four data points, the seemingly decreasing trend might be fortuitous. In fact, the same trend does not hold for upscaled models (bottom panel). The rotation of the indirect Hamiltonian relative to the reference small basis might help, especially in the case of upscaled predictions, to reduce the effect of having operators computed in a fixed (and different) basis.

We further analyze the effective Hamiltonians in terms of the decay of their matrix elements as a function of the distance between the two interacting atoms. In Figure 5, we examine

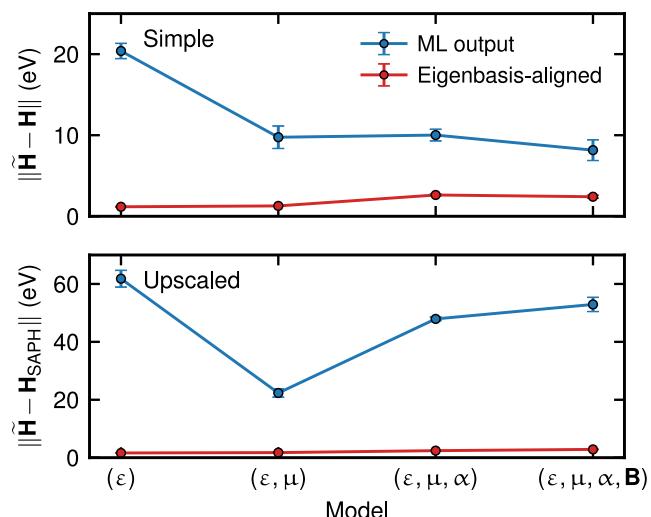


Figure 4. Norm of the deviations of the predicted Hamiltonians from their respective reference across the QM7 test set for simple indirect models (top) and upscaled indirect models (bottom). Blue lines show the deviations of ML predictions, while red lines show the deviations of eigenbasis-aligned predictions. Markers represented the values averaged over the data set and three random train/test splits, and error bars indicate the standard deviations across the three splits of the data set.

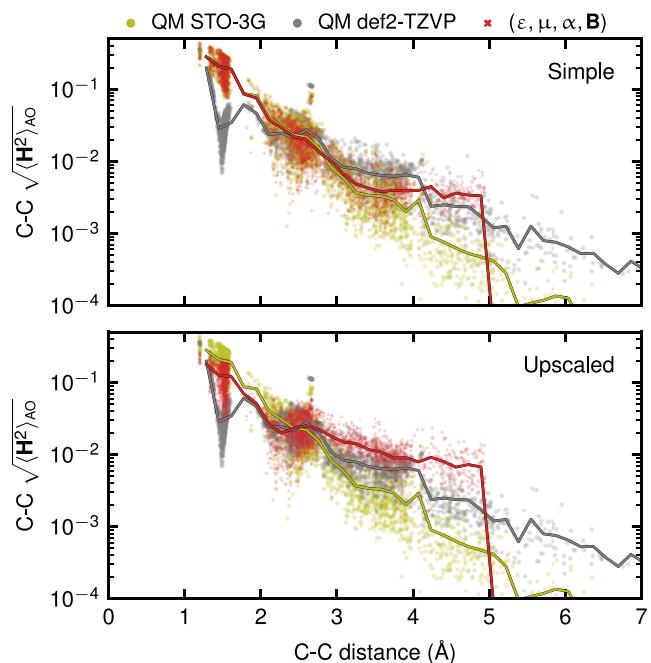


Figure 5. Decay of the carbon–carbon interaction terms in the QM9 test set as predicted from (top panel) simple indirect (targeting STO-3G observables) and (bottom panel) upscaled indirect (targeting def2-TZVP observables) models, both trained with models where EB coincides with the minimal basis. When upscaling the model to predict large-basis properties, the effective Hamiltonian adapts its behavior to match the natural scale of the reference-calculation Hamiltonian. Solid lines are moving averages of the data points.

the carbon–carbon interactions more closely. The top panel shows how $\tilde{\mathbf{H}}$ from a simple indirect model ($\varepsilon, \mu, \alpha, \mathbf{B}$), denoted by the red dots, adapts to the reference STO-3G matrix elements, while in the bottom panel, the $\tilde{\mathbf{H}}$ from the corresponding upscaled indirect model adapts to the scale and

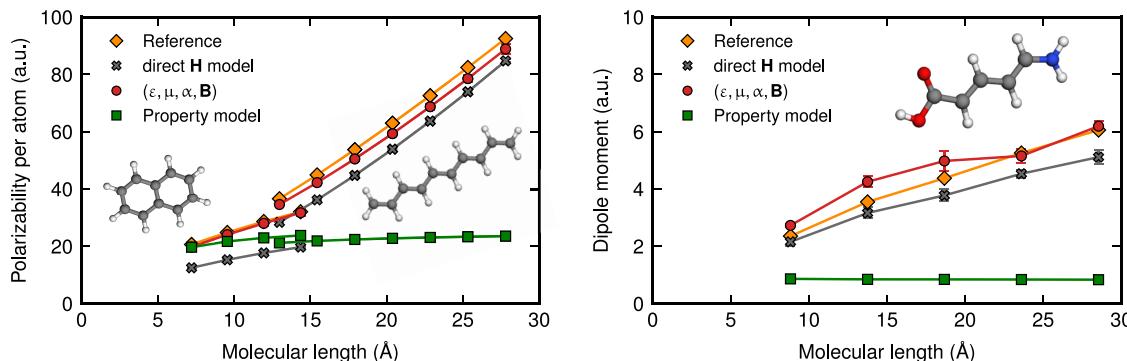


Figure 6. Performance of different ML models in predicting dielectric response properties in the extrapolative regime. (left) Norm of the polarizability per atom for increasing chain lengths of polyalkenes and polyacenes. (right) Norm of the dipole moment for a series of polyenoic amino acids with increasing chain length. In both panels, orange diamonds indicate the reference values computed with the def2-TZVP basis, gray crosses show predictions from our Hamiltonian model before fine-tuning (direct H model), red circles represent predictions from the upscaled indirect Hamiltonian model ($\epsilon, \mu, \alpha, \mathbf{B}$), and green squares correspond to the property-specific model (S) resembling AlphaML and MuML. Error bars indicate standard deviations over three independent random test/train splits. Even the direct Hamiltonian model qualitatively captures the correct scaling with system size, while the property-specific model does not. After fine-tuning, the indirect Hamiltonian model predictions become quantitatively accurate. An example molecular structure from the data set is shown in the insets.

decay characteristics of the reference def2-TZVP matrix, despite being represented in a smaller basis. This highlights that the indirect Hamiltonian models implicitly capture characteristic spatial patterns and magnitudes of Hamiltonian elements specific to the target basis, providing physically consistent effective Hamiltonians, even when trained indirectly on derived properties. The drop in the carbon–carbon interaction beyond 5 Å is a consequence of the short-range nature of the descriptors $\xi_{A_i}^{\alpha\beta\mu}$, which exclude interactions between atoms separated by more than the cutoff distance by design. However, this does not restrict the model from recovering the correct long-range behavior of observables, such as dipole moments and polarizabilities, as it is trained to indirectly reproduce these quantities, which inherently include nonlocal information.

3.4. Extrapolative Case Studies. In the extrapolative tests on QM9, shown in Figure 3, the degradation in model accuracy suffered by the Hamiltonian-based models is similar to that suffered by property models that directly estimate the dipole moment and polarizability. Refs 10,12, however, discuss specific test cases that lead to *qualitative* failures in the property models, linked to their local, atom-centered nature. The first of such test data sets involves the polarizability of long-chain polyalkenes and polyacenes, while the second consists of a series of polyenoic amino acids featuring an amine and a carboxylic acid group linked by a polyacetylene backbone. In the first series, the polarizability per atom increases with chain length owing to the progressive reduction of HOMO–LUMO gap and the resulting delocalization of electrons. Similarly, in the second data set, the molecular dipole moments grow with system size as a result of the delocalization of charges over the entire molecule. Traditional surrogate property models, which map molecular geometry to physical observables via eq 5, fail to reproduce the correct scaling behavior due to their strictly local construct. These models describe atomic environments up to distances much smaller than the characteristic length scales at which these phenomena occur, and the training set contains only small molecules.

In contrast, model architectures incorporating an effective Hamiltonian replicate the correct physical behavior. As shown in Figure 6, even a direct H model fitted to $\mathbf{H}_{\text{STO-3G}}$, that we

use as an initialization for all other models, qualitatively captures the scaling laws of the properties with respect to molecular length, and quantitative accuracy is achieved by further fine-tuning in the indirect model. Even though the Hamiltonian matrix elements are also predicted with a strictly local model, the physics-based manipulations, most importantly the diagonalization of the effective Hamiltonian, introduce the nonlocal physics that is necessary to reproduce the correct trends. These observations concretely exemplify the superior transferability of indirect Hamiltonian models compared to property-specific models.

4. CONDENSED-PHASE HAMILTONIANS

Even though we have exclusively presented molecular examples thus far, the indirect learning strategy presented in this work can also be readily applied to the condensed phase. Extended systems are routinely treated under the Born–Von Kármán periodic boundary conditions (PBC), repeating the crystal unit cell several times along each lattice vector and finally imposing the periodicity of the electronic wave function on this large (super)cell. Although accounting for periodicity is more natural and convenient in a reciprocal space representation, the structural input features parametrizing the Hamiltonian describe real-space geometries. Thus, the predicted real-space Hamiltonians corresponding to each translation \mathbf{R}_t of the unit cell can be transformed to the reciprocal space through the Bloch summation,

$$\mathbf{H}(\mathbf{k}) = \sum_t \mathbf{H}(t) e^{i\mathbf{k} \cdot \mathbf{R}_t}. \quad (12)$$

\mathbf{R}_t denotes the Bravais lattice vector and is labeled by the three integers $t = (t_x, t_y, t_z)$, indicating the number of unit-cell repetitions in each Cartesian direction. $\mathbf{H}(\mathbf{k})$ is then diagonalized to obtain a set of band energies, $\{\varepsilon_{nk}\}$, labeled by a band index n and a Brillouin zone vector, \mathbf{k} , that can be indirectly learned on an upscaled basis similar to the molecular case. As a simple example, we consider a small data set of 23 structures subsampled from the graphene data set from ref 32, which is obtained from ab initio molecular dynamics simulations at 300 K. Instead of directly learning the large basis Hamiltonian (in this case, the reference calculations are

in the DZVP (double- ζ valence polarized) basis), we construct a (ϵ) model trained on the band energies up to a few eV above the Fermi level (more specifically, all of the occupied states plus one empty state per C atom), using a calculation in the minimal SZV (single- ζ valence) basis as baseline. For these baseline density functional theory (DFT) calculations, we use a Gödecker–Teter–Hutter (GTH) pseudopotential,⁶⁵ and note that the basis functions are more spatially delocalized than those used for all-electron calculations. To facilitate modeling with (relatively) local descriptors, we target the difference between SZV and DZVP results rather than pretraining the model on the self-consistent Hamiltonians in the smaller basis as in the molecular case. The ML model predicts the correction to SZV calculations necessary to achieve large-basis accuracy at inference. The model achieves an MAE of 252 meV on the occupied band energies. The predicted bands for a hold-out test structure (a simple graphene unit cell) are compared against the reference in Figure 7.

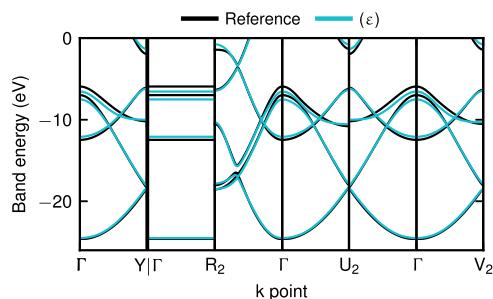


Figure 7. Comparison of reference and predicted bands for a hold-out structure from the data set. The predictions are from an upscaled SVP-basis model targeting the correction to the (ϵ) bands computed with a more converged DZVP basis.

5. CONCLUSIONS

Machine learning has been transformative for chemical sciences, offering faster and more accurate predictions of molecular properties and significantly expanding the scope of computational chemistry. In most applications, machine learning models are used as surrogates for quantum mechanical calculations. Over the past few years, there has been increasing interest in directly modeling elements of a QM calculation, such as effective single-particle Hamiltonians, from which molecular properties can be subsequently derived. Previous works in this context have focused on targeting Hamiltonian matrix representations on a specific basis or learning a reduced effective representation that reproduces observables from a calculation performed on another, larger basis.^{31–34,43}

Whenever the Hamiltonian is not an explicit target, the way it is parametrized and the targets chosen for training become part of the model architecture, introducing quantum chemistry considerations into the modeling design space. To facilitate the exploration of these possibilities, we introduced a framework that seamlessly integrates the ML predictions of effective one-electron Hamiltonians with PySCFAD, which is a differentiable quantum chemistry code. This allows us to move beyond targeting the matrix elements in a fixed basis and instead treat the Hamiltonian as an intermediate model layer that can be optimized on multiple observables computed from it through differentiable analytical operations.

A key advantage of this framework is its flexibility, allowing it to simultaneously target numerous observables, which need not be limited to those derived from a QM calculation performed on the same basis as the one used to represent the implicit model Hamiltonian. We demonstrated the capabilities of the framework by predicting molecular orbital energies, dipole moments, and polarizabilities for subsets of the QM7 and QM9 data sets, computed using both STO-3G and def2-TZVP basis sets, while restricting the model Hamiltonian to have the size and symmetries compatible with STO-3G. In both cases, we gradually constrain the model on energies, dipoles, polarizabilities, and Mayer bond order. The addition of each constraint improves the prediction of the corresponding property at the expense of a potential decrease in accuracy for the remaining ones, as the model expressivity and available training information are redistributed among the different targets. During training, the intermediate Hamiltonian is refined, starting from values obtained by prefitting against reference calculations on an STO-3G basis, to a form that more faithfully reproduces desired targets. We observe that when targeting properties computed on the def2-TZVP basis, the implicitly learned Hamiltonian tends to adapt to the slower decay of the large basis. The indirect prediction of quantum mechanical properties shows good accuracy and transferability to larger molecules despite the fact that we restricted the structural representation of the input molecules to fixed, low-body-order descriptors and their mapping to matrix elements to a linear form. For structural properties such as polarization and polarizability, the Hamiltonian-based predictions are comparable to those of bespoke models targeting the final quantity directly. Contrary to the latter, our approach correctly captures the qualitative trends in predictions for the response properties of complex molecules, such as polyenes, polyacenes, and polyenoic amino acids, where traditional property-specific models have been shown to struggle or fail entirely as they do not correctly account for electron delocalization. This approach extends beyond molecular systems as exemplified by the predictions of the DZVP band structure of graphene from a model implicitly restricted to an SZV representation of the Hamiltonian that are within 3% of the reference.

This work underscores the potential of a hybrid QM-ML paradigm in enhancing the transferability of predictive models across geometric complexity, QM details (basis set and level of theory), and physical observables. Although we restricted ourselves to simple ML models to emphasize the QM design choices, the framework we propose here is equally applicable to more sophisticated ML architectures, which could further improve the model accuracy and propel predictive modeling in complex systems. The elements of the architecture that are most closely related to the QM workflow, most notably the size and assumed symmetry of the basis, are equally important. Especially when targeting properties from a larger basis, the constraints or observables that are explicitly included in the training affect (but do not entirely determine) the alignment between the effective basis predicted by the model and that used to represent the operators that underlie those properties. For example, the use of a minimal basis representation is only partially compensated for by the ability of the model to learn an effective description of the more converged basis. In other words, restrictions on the intermediate model basis can reduce the expressive power of the model, especially for excited states and properties, such as the polarizability, that are strongly dependent on them. Overall, we suggest that when applying

ML to quantum chemical calculations, one has to take a holistic approach in which geometric features, model architecture, and QM approximations are optimized together. A higher degree of integration between the ML and QM software stack, as we implement here exploiting the autodifferentiability of PySCFAD, helps explore the design space of these hybrid models, exploit synergies between the different parts of the calculation, and ultimately improve the accuracy, transferability, and computational efficiency of the predictions.

■ ASSOCIATED CONTENT

Data Availability Statement

All the software used in this study is publicly available at ref 66. Data sets comprising molecular configurations and reference electronic structure properties can be obtained from an open repository at ref 67. An example demonstrating the use of this framework on a simplified molecular data set is available at <https://atomistic-cookbook.org/examples/hamiltonian-qm7/> and a tutorial on condensed phase Hamiltonian learning is also available at <https://atomistic-cookbook.org/examples/periodic-hamiltonian/periodic-hamiltonian.html>.

■ Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jctc.Sc00522>.

Generation of the data set and hyperparameters for computing the descriptors, as well as tabulated results for some of the figures in the main text ([PDF](#))

■ AUTHOR INFORMATION

Corresponding Authors

Jigyasa Nigam – Laboratory of Computational Science and Modeling, Institut des Matériaux, 1015 Lausanne, Switzerland; [ORCID iD](https://orcid.org/0000-0001-6857-4332); Email: jnigam@mit.edu

Michele Ceriotti – Laboratory of Computational Science and Modeling, Institut des Matériaux, 1015 Lausanne, Switzerland; Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States; [ORCID iD](https://orcid.org/0000-0003-2571-2832); Email: michele.ceriotti@epfl.ch

Authors

Divya Suman – Laboratory of Computational Science and Modeling, Institut des Matériaux, 1015 Lausanne, Switzerland; [ORCID iD](https://orcid.org/0009-0009-1483-8959)

Sandra Saade – Laboratory of Computational Science and Modeling, Institut des Matériaux, 1015 Lausanne, Switzerland; [ORCID iD](https://orcid.org/0009-0008-6163-0758)

Paolo Pegolo – Laboratory of Computational Science and Modeling, Institut des Matériaux, 1015 Lausanne, Switzerland; [ORCID iD](https://orcid.org/0000-0003-1491-8229)

Hanna Türk – Laboratory of Computational Science and Modeling, Institut des Matériaux, 1015 Lausanne, Switzerland

Xing Zhang – Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States

Garnet Kin-Lic Chan – Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena,

California 91125, United States; [ORCID iD](https://orcid.org/0000-0001-8009-6038)

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jctc.Sc00522>

Author Contributions

#D.S. and J.N. contributed equally to this work. M.C. and G.K.C. designed the research strategy; D.S., J.N., P.P., and S.S. performed simulations and prepared figures; J.N. and P.P. developed the ML software infrastructure, with contributions from D.S., S.S., and H.T.; X.Z. and G.K.C. developed the PySCFAD interface; J.N., P.P., and D.S. wrote the first version of the article; all authors commented on the manuscript and contributed to the final version.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

M.C., D.S., J.N., and S.S. acknowledge funding from the European Research Council (ERC) under the research and innovation program (Grant Agreement No. 101001890-FIAMMA), an industrial grant from Samsung, and the NCCR MARVEL, funded by the Swiss National Science Foundation (SNSF, Grant Number 205602). P.P. acknowledges funding from Samsung. H.T. acknowledges funding from the Walter Benjamin program (project number 534232519) of the German Research Foundation (DFG). G.K.C. and X.Z. acknowledge support from the US Department of Energy, Office of Science, Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division under Triad National Security, LLC (“Triad”) contract Grant 89233218CNA000001.

■ REFERENCES

- (1) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, No. 146401.
- (2) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, No. 136403.
- (3) Thompson, A. P.; Swiler, L. P.; Trott, C. R.; Foiles, S. M.; Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **2015**, *285*, 316–330.
- (4) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Schmidt, T.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, No. 2453.
- (5) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csanyi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds.; Curran Associates, Inc., 2022; Vol. 35, pp 11423–11436.
- (6) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **2009**, *131*, No. 124101.
- (7) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, No. 058301.
- (8) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **2019**, *99*, No. 014104.
- (9) Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **2021**, *121*, 10037–10072.

- (10) Veit, M.; Wilkins, D. M.; Yang, Y.; DiStasio, R. A.; Ceriotti, M. Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles. *J. Chem. Phys.* **2020**, *153*, No. 024113.
- (11) Sun, J.; Cheng, L.; Miller, T. F., III Molecular dipole moment learning via rotationally equivariant derivative kernels in molecular-orbital-based machine learning. *J. Chem. Phys.* **2022**, *157*, No. 104109.
- (12) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 3401–3406.
- (13) Zhang, L.; Chen, M.; Wu, X.; Wang, H.; Weinan, E.; Car, R. Deep neural network for the dielectric response of insulators. *Phys. Rev. B* **2020**, *102*, No. 041121.
- (14) Zhang, Y.; Jiang, B. Universal machine learning for the response of atomistic systems to external fields. *Nat. Commun.* **2023**, *14*, No. 6424.
- (15) Feng, C.; Xi, J.; Zhang, Y.; Jiang, B.; Zhou, Y. Accurate and interpretable dipole interaction model-based machine learning for molecular polarizability. *J. Chem. Theory Comput.* **2023**, *19*, 1207–1217.
- (16) Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **2019**, *5*, 57–64.
- (17) Lewis, A. M.; Grisafi, A.; Ceriotti, M.; Rossi, M. Learning electron densities in the condensed phase. *J. Chem. Theory Comput.* **2021**, *17*, 7203–7214.
- (18) Fabrizio, A.; Briling, K. R.; Girardier, D. D.; Corminboeuf, C. Learning on-top: Regressing the on-top pair density for real-space visualization of electron correlation. *J. Chem. Phys.* **2020**, *153*, No. 204111.
- (19) Rackers, J. A.; Tecot, L.; Geiger, M.; Smidt, T. E. A recipe for cracking the quantum scaling limit with machine learned electron densities. *Mach. Learn.: Sci. Technol.* **2023**, *4*, No. 015027.
- (20) Koker, T.; Quigley, K.; Taw, E.; Tibbetts, K.; Li, L. Higher-order equivariant neural networks for charge density prediction in materials. *npj Comput. Mater.* **2024**, *10*, No. 161.
- (21) Li, C.; Sharir, O.; Yuan, S.; Chan, G. K. Image super-resolution inspired electron density prediction. 2024. arXiv:2402.12335. arXiv.org e-Print archive. <https://arxiv.org/abs/2402.12335>.
- (22) Shao, X.; Paetow, L.; Tuckerman, M. E.; Pavanello, M. Machine learning electronic structure methods based on the one-electron reduced density matrix. *Nat. Commun.* **2023**, *14*, No. 6281.
- (23) Febrer, P.; Jørgensen, P.; Pruneda, M.; Garcia, A.; Ordejón, P.; Bhowmik, A. Graph2mat: Universal graph to matrix conversion for electron density prediction. *Mach. Learn.: Sci. Technol.* **2024**, *6*, No. 025013.
- (24) Zhang, L.; Mazzeo, P.; Nottoli, M.; Cignoni, E.; Cupellini, L.; Stamm, B. A symmetry-preserving and transferable representation for learning the kohn-sham density matrix. 2025. arXiv:2503.08400. arXiv.org e-Print archive. <https://arxiv.org/abs/2503.08400>.
- (25) Pfau, D.; Spencer, J. S.; Matthews, A. G. D. G.; Matthew, W.; Foulkes, C. Ab initio solution of the many-electron schrödinger equation with deep neural networks. *Phys. Rev. Res.* **2020**, *2*, No. 033429.
- (26) Schätzle, Z.; Szabó, P. B.; Mezera, M.; Hermann, J.; Noé, F. Deepqmc: An open-source software suite for variational optimization of deep-learning molecular wave functions. *J. Chem. Phys.* **2023**, *159*, No. 094108.
- (27) Hermann, J.; Spencer, J.; Choo, K.; Mezzacapo, A.; Matthew, W.; Foulkes, C.; Pfau, D.; Carleo, G.; Noé, F. Ab initio quantum chemistry with neural-network wavefunctions. *Nat. Rev. Chem.* **2023**, *7*, 692–709.
- (28) Pescia, G.; Nys, J.; Kim, J.; Lovato, A.; Carleo, G. Message-passing neural quantum states for the homogeneous electron gas. *Phys. Rev. B* **2024**, *110*, No. 035108.
- (29) Scherbela, M.; Gerard, L.; Grohs, P. Towards a transferable fermionic neural wavefunction for molecules. *Nat. Commun.* **2024**, *15*, No. 120.
- (30) Gerard, L.; Scherbela, M.; Sutterud, H.; Foulkes, M.; Grohs, P. Transferable neural wavefunctions for solids. 2024. arXiv:2405.07599. arXiv.org e-Print archive. <https://arxiv.org/abs/2405.07599>.
- (31) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **2019**, *10*, No. 5024.
- (32) Li, H.; Wang, Z.; Zou, N.; Ye, M.; Xu, R.; Gong, X.; Duan, W.; Xu, Y. Deep-learning density functional theory hamiltonian for efficient ab initio electronic-structure calculation. *Nat. Comput. Sci.* **2022**, *2*, 367–377.
- (33) Gong, X.; Li, H.; Zou, N.; Xu, R.; Duan, X.; Xu, Y. General framework for $e(3)$ -equivariant neural network representation of density functional theory hamiltonian. *Nat. Commun.* **2023**, *14*, No. 2848.
- (34) Zhang, L.; Onat, B.; Dusson, G.; McSloy, A.; Anand, G.; Maurer, R. J.; Ortner, C.; Kermode, J. R. Equivariant analytical mapping of first principles hamiltonians to accurate and transferable materials models. *npj Comput. Mater.* **2022**, *8*, No. 158.
- (35) Nigam, J.; Willatt, M. J.; Ceriotti, M. Equivariant representations for molecular Hamiltonians and N-center atomic-scale properties. *J. Chem. Phys.* **2022**, *156*, No. 014115.
- (36) Hegde, G.; Bowen, R. C. Machine-learned approximations to density functional theory hamiltonians. *Sci. Rep.* **2017**, *7*, No. 42669.
- (37) Westermayr, J.; Maurer, R. J. Physically inspired deep learning of molecular excitations and photoemission spectra. *Chem. Sci.* **2021**, *12*, 10755–10764.
- (38) Unke, O. T.; Bogojeski, M.; Gastegger, M.; Geiger, M.; Smidt, T.; Müller, K.-R. In *Se(3)-Equivariant Prediction of Molecular Wavefunctions and Electronic Densities*, NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems; NeurIPS, 2024.
- (39) Yin, S.; Pan, X.; Zhu, X.; Gao, T.; Zhang, H.; Wu, F.; He, L. Towards harmonization of $so(3)$ -equivariance and expressiveness: a hybrid deep learning framework for electronic-structure hamiltonian prediction. *Mach. Learn.: Sci. Technol.* **2024**, *5*, No. 045038.
- (40) Wang, Y.; Li, H.; Tang, Z.; Tao, H.; Wang, Y.; Yuan, Z.; Chen, Z.; Duan, W.; Xu, Y. Deep-2: Enhancing deep-learning electronic structure via an equivariant local-coordinate transformer. 2024. arXiv:2401.17015. arXiv.org e-Print archive. <https://arxiv.org/abs/2401.17015>.
- (41) Zhong, Y.; Yu, H.; Su, M.; Gong, X.; Xiang, H. Transferable equivariant graph neural networks for the hamiltonians of molecules and solids. *npj Comput. Mater.* **2023**, *9*, No. 182.
- (42) Zhong, Y.; Yang, J.; Xiang, H.; Gong, X.; et al. Universal machine learning kohn-sham hamiltonian for materials. *Chin. Phys. Lett.* **2024**, *41*, No. 077103.
- (43) Cignoni, E.; Suman, D.; Nigam, J.; Cupellini, L.; Mennucci, B.; Ceriotti, M. Electronic excited states from physically constrained machine learning. *ACS Cent. Sci.* **2024**, *10*, 637–648.
- (44) Tang, H.; Xiao, B.; He, W.; Subasic, P.; Harutyunyan, A. R.; Wang, Y.; Liu, F.; Xu, H.; Li, J. Approaching coupled-cluster accuracy for molecular electronic structures with multi-task learning. *Nat. Comput. Sci.* **2025**, *5*, 144–154.
- (45) Tamayo-Mendoza, T.; Kreisbeck, C.; Lindh, R.; Aspuru-Guzik, A. Automatic differentiation in quantum chemistry with applications to fully variational hartree-fock. *ACS Cent. Sci.* **2018**, *4*, 559–566.
- (46) Li, H.; Collins, C.; Tanha, M.; Gordon, G. J.; Yaron, D. J. A density functional tight binding layer for deep learning of chemical hamiltonians. *J. Chem. Theory Comput.* **2018**, *14*, 5764–5776.
- (47) Tan, C.; Wei, J.; Pickard, C. J.; Witt, W. C. Automatic differentiation for orbital-free density functional theory. *J. Chem. Phys.* **2023**, *158*, No. 124801.
- (48) Stishenko, P.; McSloy, A.; Onat, B.; Hourahine, B.; Maurer, R. J.; Kermode, J. R.; Logsdail, A. Integrated workflows and interfaces for data-driven semi-empirical electronic structure calculations. *J. Chem. Phys.* **2024**, *161*, No. 012502.

- (49) Zhangand, X.; Chan, G. K.-L. Differentiable quantum chemistry with pyscf for molecules and materials at the mean-field level and beyond. *J. Chem. Phys.* **2022**, *157*, No. 204801.
- (50) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, No. 058301.
- (51) Blum, L. C.; Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.
- (52) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, No. 095003.
- (53) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, No. 140022.
- (54) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.
- (55) Nigam, J.; Pozdnyakov, S.; Ceriotti, M. Recursive evaluation and iterative contraction of N -body equivariant features. *J. Chem. Phys.* **2020**, *153*, No. 121101.
- (56) van der Oord, C.; Dusson, G.; Csányi, G.; Ortner, C. Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials. *Mach. Learn.: Sci. Technol.* **2020**, *1*, No. 015004.
- (57) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, No. 036002.
- (58) Fabrizio, A.; Briling, K. R.; Corminboeuf, C. SPAHM: the spectrum of approximated hamiltonian matrices representations. *Digital Discovery* **2022**, *1*, 286–294.
- (59) Welborn, M.; Cheng, L.; Miller, T. F., III Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772–4779.
- (60) Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F., III Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **2020**, *153*, No. 124111.
- (61) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **2018**, *148*, No. 241730.
- (62) Bartók, A. P.; Kondor, Risi.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, No. 184115.
- (63) Bridgeman, A. J.; Cavigliasso, G.; Ireland, L. R.; Rothery, J. The mayer bond order as a tool in inorganic chemistry. *J. Chem. Soc., Dalton Trans.* **2001**, 2095–2108.
- (64) Baroni, S.; de Gironcoli, S.; Corso, A. D.; Giannozzi, P. Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **2001**, *73*, No. 515.
- (65) Goedecker, S.; Teter, M.; Hutter, J. Separable dual-space gaussian pseudopotentials. *Phys. Rev. B* **1996**, *54*, No. 1703.
- (66) Nigam, J.; Pegolo, P.; Ceriotti, M. Integrating ML for Hamiltonian and electronic structure 2001 <https://github.com/curiosityS4/mleeca>.
- (67) Suman, D.; Nigam, J.; Saade, S.; Pegolo, P.; Tuerk, H.; Zhang, X.; Chan, G. K.-L.; Ceriotti, M. Exploring the design space of machine-learning models for quantum chemistry with a fully differentiable framework. 2025. arXiv:2504.01187. arXiv.org e-Print archive. <https://arxiv.org/abs/2504.01187>.



CAS BIOFINDER DISCOVERY PLATFORM™

BRIDGE BIOLOGY AND CHEMISTRY FOR FASTER ANSWERS

Analyze target relationships,
compound effects, and disease
pathways

Explore the platform

