Article:
Shah, A.A., Leung, P.K. and Xing, W.W. (2025) Rapid high-fidelity quantum simulations using multi-step nonlinear autoregression and graph embeddings. npj Computational Materials, 11 (1). 57. ISSN: 2057-3960

https://doi.org/10.1038/s41524-024-01479-0

npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01479-0

Rapid high-fidelity quantum simulations using multi-step nonlinear autoregression and graph embeddings

Akeel A. Shah [1] ✉, P. K. Leung[1] & W. W. Xing[2] ✉

The design and high-throughput screening of materials using machine-learning assisted quantum-mechanical simulations typically requires the existence of a very large data set, often generated from simulations at a high level of theory or fidelity. A single simulation at high fidelity can take on the order of days for a complex molecule. Thus, although machine learning surrogate simulations seem promising at first glance, generation of the training data can defeat the original purpose. For this reason, the use of machine learning to screen or design materials remains elusive for many important applications. In this paper we introduce a new multi-fidelity approach based on a dual graph embedding to extract features that are placed inside a nonlinear multi-step autoregressive model. Experiments on five benchmark problems, with 14 different quantities and 27 different levels of theory, demonstrate the generalizability and high accuracy of the approach. It typically requires a few 10s to a few 1000's of high-fidelity training points, which is several orders of magnitude lower than direct ML methods, and can be up to two orders of magnitude lower than other multi-fidelity methods. Furthermore, we develop a new benchmark data set for 860 benzoquinone molecules with up to 14 atoms, containing energy, HOMO, LUMO and dipole moment values at four levels of theory, up to coupled cluster with singles and doubles.

It has long been the goal to use quantum-mechanical (QM) simulations for high-throughput materials screening and design[1]. This goal, however, remains elusive given the prohibitive computational costs associated with approaches that involve high levels of theory (high-fidelity). A high-fidelity is essential for many applications; for example, semiconductor applications require accurate calculations of electronic band structures and energy levels, energy storage and conversion applications require precise predictions of the energetics of charge transfer processes, and pharmaceutical applications require accurate predictions of binding energies and vibrational frequencies.

At the lowest fidelity, semi-empirical methods[2] and the Hartree-Fock (HF) method[3] have attractive scalings (in practise) of $O(N^3)$ or lower, where N is the number of electrons, the number of basis functions or some other suitably-defined system size. Various post-HF approaches have been developed to account for the neglected electron correlation effects in these methods. The configuration interaction (CI) method[4] includes excitations *via* determinants, leading to scalings of $O(N^6)$ when single and double excitations are included (the CISD method). Møller-Plesset (MPn) theory[5] expands the wavefunction and energy as asymptotic series (with n terms) about the HF values, with a minimum scaling of $O(N^5)$ in the case of MP2. Coupled cluster (CC) theory[6] expands the wavefunction using an exponential excitation operator that incorporates higher-level excitations, with a minimum scaling of $O(N^6)$, namely for single and double excitations (CCSD). Incorporating triple excitations (CCSDT) has $O(N^8)$ cost, while an approximate treatment using perturbation theory (CCSD(T)) scales as $O(N^7)$. Greens' function (GW) methods[7], commonly used for solid-state materials, use a self-energy operator to improve upon the electronic energies and wave functions obtained from HF, increasing the cost to at least $O(N^4)$.

Kohn-Sham (KS) density functional theory (DFT)[8] strikes perhaps the best balance between accuracy and computational costs, which are in the range $O(N^3)$ to $O(N^4)$. It reformulates the original problem as a minimization of the total energy (as a functional of the electron density)[9], placing much of the uncertainty inside an exchange-correlation (XC) energy functional contribution[10–12]. A hierarchy of XC functionals with increasing sophistication leads to different levels of accuracy. Despite its many successes, however, DFT is insufficiently accurate for a wide range of applications. To obtain more accurate results, composite methods such as G4MP2[13] are often employed. Such methods combine post-HF methods with a basis set extrapolation and additional thermal corrections to efficiently enhance

[1]School of Energy and Power Engineering, Chongqing University, Shapingba, Chongqing, China. [2]School of Mathematics and Statistics, University of Sheffield, Sheffield, UK. ✉e-mail: ashah@cqu.edu.cn; w.xing@sheffield.ac.uk

accuracy, but with scalings of $O(N^6)$ to $O(N^7)$ they can only be applied to small systems. The basis set also influences the accuracy and cost. Minimal basis sets such as STO-3G[14] are computationally efficient, but for higher accuracy split valence sets such as 6-31G*[15] and larger localized sets such as cc-pVDZ[16] and their augmented equivalents are required.

In an attempt to circumvent the slow computational times related to QM computations, various machine learning (ML) approaches have been adopted[17–19]. The most obvious approach is to learn a map directly between the system (molecules) and chosen outputs[20,21]. Other approaches include data-driven XC functionals[22], mappings between the electron density and functional contributions to the energy, and mappings between the external potential and density of states[23,24]. Arguably, the main challenge in a ML approach lies in finding descriptors (numerical encodings of molecules) that exhibit a good causal link to the target[18,25]. This has been a longstanding problem in the related area of quantitative structure-activity relationship modeling[25]. Conventional descriptors such as bags of bonds (BoB)[26] and fingerprints[27] have a limited capacity to accurately describe molecules. For example, Faber et al.[28] found that 100k training samples were insufficient to reach a desired accuracy for 6 out of 9 selected targets, regardless of the ML method or descriptor. Localized descriptors such as the smooth overlap of atomic positions (SOAP)[29] and atom-centered symmetry functions (ACSF)[30] encode atomic environments using basis function expansions. These local expansions can then be combined to define a global descriptor. In some cases they can lead to higher accuracies than conventional global descriptors[31,32], although not consistently[33,34].

In certain applications, small[18] or relatively small[31,35] data sets can yield acceptable accuracies. In general, however, a large data set is required[28,36]. Generating the training set, especially for high-fidelity (hi-fi) predictions, can thus defeat the original purpose, unless perhaps existing databases are available. Non-trivial strategies for facilitating the use of small data sets are active learning, transfer learning and semi-supervised learning. Recent reviews of small-data approaches can be found in refs. 35 and [37]. Active learning, typically based on entropy or variance reduction, consists of parsimoniously selecting inputs for full simulations in order to generate a data set. It is naturally employed in database construction or in an optimization strategy with a specific objective in mind.

Transfer learning[38] involves the (pre-)training of models on surplus data sets, before transferring features extracted from the best-performing models to a separate model of the target data set. There are several drawbacks to this approach. Firstly, it involves pre-training a large number of models, typically on the order of 1000, on at least one very large data set (usually several). The data sets must be compatible, meaning similar materials, similar target quantities and similar generating methods, so that there are strong correlations between the outputs. For most new applications, such a large volume of compatible data is unlikely to exist. In almost all cases, the increase in accuracy is negligible[39,40] or modest[38,40,41], and the final model is not guaranteed to yield accurate predictions[38,42]. Thus, the prospective gains in accuracy do not necessarily justify the substantial resources and time required for data acquisition, data cleaning, pre-training and construction of the final model, especially given that few of these steps can be automated. Semi-supervised learning can be used to artificially increase the size of the training set. The classic approach trains a model on the available data and then applies the trained model to unlabeled data in order to generate predictions. The predicted outputs are added to the original training set and the model is re-trained with the augmented data. Studies have demonstrated that such a strategy can reduce the error on small data sets by around a quarter to a third at most[43,44], which is not especially significant in many contexts.

State-of the-art graph neural networks[19] such as MatErials Graph Network (MEGNET)[45], SchNet[46] and DimeNet[47,48] use a message passing mechanism to embed (featurize) molecular structures in a graph, potentially providing better encodings of molecular systems. Although they improve accuracy, numerous studies have shown that they still require $O(10k) - O(100k)$ training points when applied to standard data sets[17,45,46,49]. Moreover, they do not consistently outperform standard techniques. For example, Fung et al.[49] compared four state-of-the art (SOTA) graph-based methods to a conventional network with a SOAP input, and found the latter to be marginally more accurate on 2 out of 5 data sets. At least several thousand and up to 10k training points were required.

The overall lack of consistency of ML algorithms in terms of the number of training points they require and their generalizability is a major hurdle to their application in materials science. As alluded to above, the major cause of the issues is the inherent limitations of descriptors as regressors, whether they are conventional, localized or graph-based. An alternative to direct ML that can potentially overcome some of its drawbacks is multi-fidelity (MF) modeling[50], which aims to predict hi-fi results by leveraging equivalent low-fidelity (lo-fi) data[51]. Due to the correlations between lo-fi and hi-fi data and a reduced reliance on a molecular descriptor, MF approaches can dramatically reduce the number of hi-fi results required to attain a given level of accuracy[52–54]. The earliest MF approach for QM calculations was the Δ-ML model of Ramakrishnan et at.[52,53], which is based on learning residuals between successive fidelities *via* kernel regression. This method was extended to the multi-level combination quantum machine learning (CQML) method[55], which introduces additional fidelities across basis sets and also approximates the lo-fi result using direct ML. Taking a different route[54], extended MEGNET to MF data (MF-MEGNET) by treating the fidelity index as an input; see also Fare et al.[56] for a similar approach using a Gaussian process (GP) model. General SOTA MF approaches include GP linear autoregression (GPLAR)[51], GP nonlinear autoregression (GPNAR)[57], stochastic collocation (SC)[58] and their variants[59]. Used GPLAR with a force field model and DFT as the low-fi and hi-fi models, respectively, to learn the lattice energies of crystals. In a similar fashion, Tran et al.[60] used GPLAR for MD simulations of ternary random alloys, with classical MD as the lo-fi model and DFT ab-initio MD as the hi-fi model.

In this paper, we present Multi-Fidelity autoregressive GP with Graph Embeddings for Molecules (MFGP-GEM), a new approach for hi-fi QM simulations. MFGP-GEM uses a two-step spectral embedding of molecules in a graph *via* manifold learning, namely diffusion maps[61,62]. The embedding is combined with data at an arbitrary number of low-medium fidelities in order to define inputs to a multi-step nonlinear autoregressive GP. The method is inspired by recent developments in graph embeddings for GP models, which are usually based on the graph Laplacian[63,64]. These embeddings can be used to define distances between irregular structures such as molecules[65]. Compared to deep networks, GPs are explainable, they generally require smaller data sets for a given accuracy, and they furnish error bounds.

MFGP-GEM is compared to SOTA ML and MF approaches, including MF-MEGNET, Δ-ML and CQML. We develop a new benchmark data set of 860 benzoquinone molecules (up to 14 atoms) comprising energy, HOMO, LUMO and dipole moment calculations at four fidelities: HF, DFT with the B3LYP (Becke, 3-parameter, Lee-Yang-Parr) functional[66], MP2 and CCSD, all employing a cc-pVDZ basis. In order to demonstrate generalizability, a further four data sets are considered, with a total of 14 different quantities and 27 different fidelities. The results reveal that MFGP-GEM typically requires a few 10s to a few 1000's of hi-fi training points to attain high accuracy, orders of magnitude lower than direct ML methods and significantly fewer than other QM MF methods. Descriptions of the QM simulation methods, data sets, ML and MF methods, and molecular descriptors considered in this paper are provided in the Supplementary, together with additional results.

Results

Data sets and an outline of MFGP-GEM

Summary statistics of the data sets for selected properties and fidelities are shown in Table 1. The first data set pertains to 860 benzoquinone molecules with between 6 and 14 atoms, examples of which are shown in Fig. 1. The data set includes energy E_0, HOMO, LUMO and dipole moment μ values using HF, DFT-B3LYP, MP2 and CCSD with a cc-pVDZ basis. Generation of this data set is described in "Generation of the Quinone data set", with

Table 1 | Statistics of the data sets for selected properties and fidelities

Data set	Property	Fidelity	Mean	Standard deviation	Number	Units
Quinone	E_0	CCSD-cc-pVDZ	−22567.92	27571.0	860	eV
Quinone	μ	MP2R-cc-pVDZ	3.36	1.64	536	D
Quinone	HOMO	DFT-B3YLP-cc-pVDZ	−6.70	0.58	860	eV
Quinone	LUMO	DFT-B3YLP-cc-pVDZ	−1.99	1.41	860	eV
QM7bE	$\lvert E_{\text{eff}}\rvert$	CCSD(T)-cc-pVDZ	16.24	20.64	7211	kcal/mol
QM7bE	$\lvert E_{\text{eff}}\rvert$	CCSD(T)-6-31G	18.38	24.48	7211	kcal/mol
QM7bE	$\lvert E_{\text{eff}}\rvert$	CCSD(T)-STO-3G	20.68	30.08	7211	kcal/mol
Alexandria	$\lvert\Delta_f H^0\rvert$	G4	251.9	360.3	2067	kJ/mol
Alexandria	C_V	G4	103.2	48.6	2067	J/mol/K
Alexandria	S^0	G4	341.9	71.6	2067	J/mol/K
Alexandria	$\lvert\Delta_f H^0\rvert$	W1BD	289.3	447.4	701	kJ/mol
Alexandria	C_V	W1BD	64.9	26.1	701	J/mol/K
Alexandria	S^0	W1BD	286.4	47.15	701	J/mol/K
Alexandria	$\lvert\Delta_f H^0\rvert$	experimental	236.7	292.6	1226	kJ/mol
Alexandria	C_V	experimental	108.8	47.9	942	J/mol/K
Alexandria	S^0	experimental	352.4	83.2	1044	J/mol/K
Alexandria	α	DFT-B3LYP-aug-cc-pVTZ	11.55	5.58	2347	Å^3
Alexandria	μ	DFT-B3LYP-aug-cc-pVTZ	1.76	1.60	2347	D
Alexandria	α	experimental	12.37	4.23	1206	Å^3
Alexandria	μ	experimental	1.93	1.27	1043	D
Bandgap	E_g	HSE	1.78	2.28	6030	eV
Bandgap	E_g	GLLB-SC	4.18	2.72	2126	eV

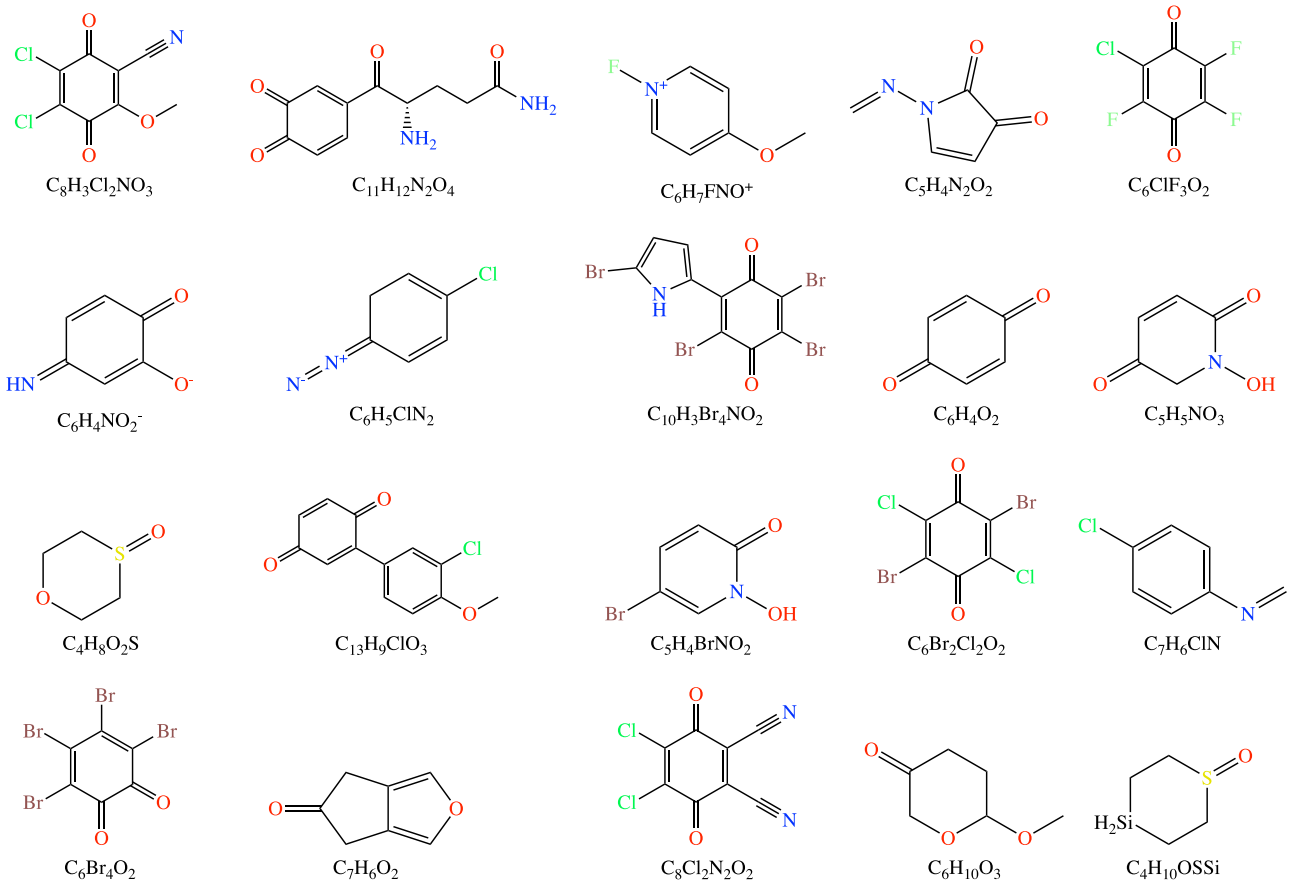

Fig. 1 | Example benzoquinones from the Quinone data set.

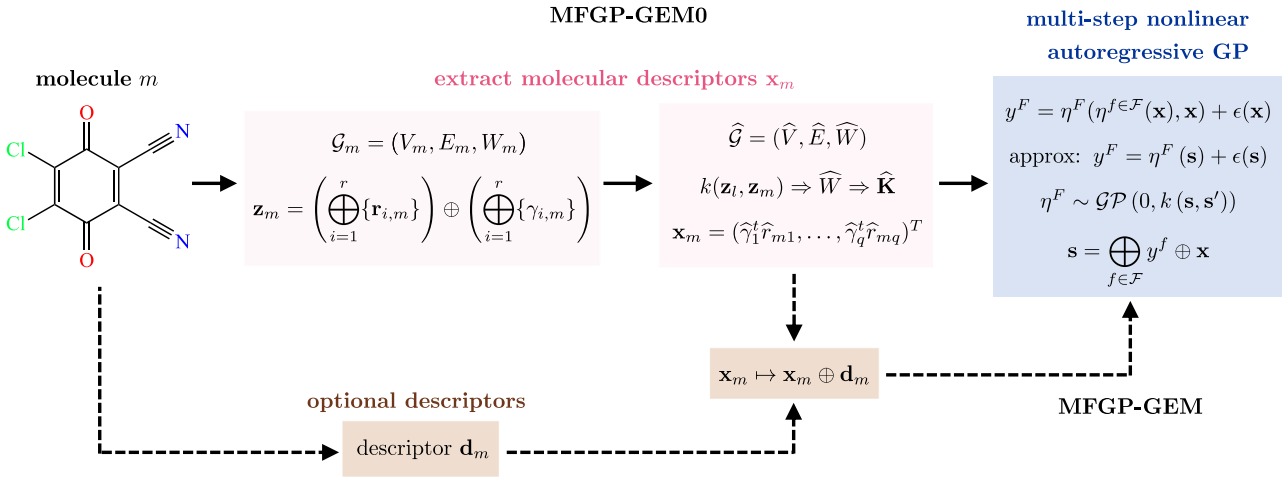

Fig. 2 | An illustration of MFGP-GEM. A descriptor $\mathbf{z}_m$ for a molecule m is generated by a diffusion map embedding on a graph $\mathcal{G}_m$ with nodes $v_{i,m} \in V_m$ identified with the atoms, and weights $w_{ij,m} \in W_m$ encoding connectivity information. Each molecule is then identified with a node $\widehat{v}_m \in \widehat{V}$ on a new graph $\widehat{\mathcal{G}}$ with weights $\widehat{w}_{mn} \in \widehat{W}$ defined *via* kernel values based on the $\mathbf{z}_m$. This defines an adjacency matrix $\widehat{\mathbf{K}}$ from which the final descriptors $\mathbf{x}_m$ are extracted using a second diffusion map embedding. The $\mathbf{x}_m$ are employed in a autoregressive GP for the hi-fi map $\eta^F(\eta^{f\in\mathcal{F}}(\mathbf{x}), \mathbf{x}) \approx y^f$, in which $\mathcal{F}$ is a subset of the fidelity indices. This leads to the method referred to as MFGP-GEM0. Other descriptors $\mathbf{d}_m$ can be combined with $\mathbf{x}_m$ ($\oplus$ denotes a concatenation), which leads to the method MFGP-GEM.

more details in Supplementary Section 2.1. We call this the Quinone data set.

The other data sets are: (1) QM7b[67], which contains results using GW, self-consistent screening (SCS)[68], DFT-PBE0[69] (a hybrid of Perdew-Burke-Ernzerhof (PBE)[11] and Hartree-Fock XC contributions) and ZINDO (a semi-empirical method) for 7211 molecules from the GDB-13 database; (2) an extended QM7b data set (QM7bE) for the 7211 molecules[55] containing calculations of the effective averaged atomization energy E_{eff} using HF, MP2, and CCSD(T) with cc-pVDZ, 6-31G and STO-3G basis sets; (3) the Alexandria library[70] comprising the enthalpy of formation Δ_f^0, specific heat capacity at constant volume C_V, absolute entropy at room temperature S^0, μ and the trace polarisability α for 2704 drug-like compounds, at HF- 6-311G**, DFT-B3LYP-aug-cc-pVTZ, G-n (Gaussian-n, n = 2, 3, 4)[71,72], Weizmann-1 Brueckner doubles (W1BD)[73] and complete basis set quadratic configuration interaction with single and double substitutions (CBS-QB3)[74,75] levels, together with some experimental data; (4) band gap values E_g of functional inorganic materials from the Materials Project, with ca. 52 k values at DFT-PBE level, 2290 with a Gritsenko-Leeuwen-Lenthe-Baerends with solid correction (GLLB-SC) functional[76], and 6030 with a Heyd-Scuseria-Ernzerhof (HSE) functional[77]. More details related to these data sets are provided in Supplementary Section 2.2.

Molecular descriptors include Coulomb matrices and their eigenvalues[78], bags of bonds (BoB)[26], Molecular ACCess System (MACCS) keys[79], sum over bonds (SoB)[80] and fingerprints[27,81], which is by no means an exhaustive list. In a comprehensive study, Elton et al.[18] identified SoB and E-state fingerprints as the most accurate descriptors. In this paper we focused on E-state and Morgan fingerprints, SoB, and MACCS keys. For the direct machine learning approaches we also investigated SOAP and ACSF inputs. Finally, reduced-dimensional representations of these descriptors derived from a principal component analysis (PCA) and a sparse PCA (SPCA)[82] were also considered. All of these methods are outlined in Supplementary Section 3. Details of their implementation are provided in "Implementation details".

Consider an F-fidelity data set, with observations y^f at fidelities $f = 1, \ldots, F$, of some quantity such as the HOMO. A statistical model of the following form is assumed: $y_m^f = \eta^f(\mathbf{x}_m) + \epsilon^f(\mathbf{x}_m)$, in which ϵ^f is an error term and $\mathbf{x}_m \in \mathbb{R}^d$, $m = 1, \ldots, M$, are vectorized molecular descriptors for molecules labeled m. The function $\eta^f(\cdot)$ denotes the true (latent) value that would be calculated in the absence of error. The aim is to approximate the map $\eta^F(\mathbf{x})$ so that hi-fi predictions can be made for an arbitrary molecule m with descriptor $\mathbf{x}$. The MF dataset can be written compactly as $\{\mathbf{X}^f\}, \{\mathbf{y}^f\}$ with the assumption that $\mathbf{X}^f \subseteq \mathbf{X}^{f-1}$, in which $\mathbf{X}_f^f \in \mathbb{R}^{d\times M_f}$ is an array of inputs $\mathbf{x}_m^f$, $m = 1, \ldots, M_f$, and $\mathbf{y}^f = (y_1^f, \ldots, y_{M_f}^f) \in \mathbb{R}^{M_f}$ is a vector of corresponding outputs at fidelity f. MFGP-GEM assumes the model $y^F = \eta^F(\eta^{f\in\mathcal{F}}(\mathbf{x}), \mathbf{x}) + \epsilon(\mathbf{x})$, in which $\mathcal{F} \subseteq \{F-1, \ldots, 1\}$ and ϵ is a generalized error, with GPs over η^F and ϵ. The descriptor $\mathbf{x}$ is obtained using a dual graph spectral embedding approach based on diffusion maps (Section "Gaussian processes based on spectral graph embeddings via diffusion maps"), optionally concatenated with additional descriptors $\mathbf{d}$ such as a MACCS key. The overall approach is illustrated in Fig. 2, in which MFGP-GEM0 refers to MFGP-GEM without additional descriptors $\mathbf{d}$.

Direct ML methods used for comparison are: classic GPs (Supplementary Section 4.2); support vector regression, convolutional networks (CNNs), multi-layer perceptrons (MLPs), and graph convolutional networks (GCNs), all of which are described in Supplementary Section 5; and MEGNET, also described in Supplementary Section 5. The MF methods used for comparison are LARGP, NARGP, SC, MF-MEGNET, CQML and Δ-ML, the first three of which are described in Supplementary Section 6. Whenever available, we used error values from the original sources.

To conserve space, the direct ML results are presented in Supplementary Section 7. These results underline the poor performance of direct ML on small or relatively small data sets. Details of the implementations of the ML and MF methods are provide in "Implementations of direct machine learning and other multi-fidelity methods". The root mean square error (RMSE), mean absolute error (MAE) and R^2 are defined in Supplementary Section 8.

Two-fidelity simulations

Quinone data set. The RMSE values on the Quinone data set using MF approaches are shown in Fig. 3. The results for MF-MEGNET and CQML are not shown since the RMSE values were at least an order of magnitude higher. For example, MF-MEGNET (CQML) yielded RMSE values of 1423 (847) eV for E_0 and 1.66 (1.64) D for μ using 80% of the data for training, with the combinations CCSD:HF for E_0 and MP2U:HF for μ. Meaningful results could not be realized for any lo-fi:hi-fi training ratio, likely to be due to the inaccuracy of the map from the input to lo-fi data. In all methods, the results are shown for PCA reduced MACCS keys (50 principal components). PCA reduced E-state fingerprints yielded the second highest accuracy, followed by PCA-SoB and PCA-Morgan. In general, PCA versions provided higher accuracies than the raw descriptors, while SPCA versions were inferior in all cases.

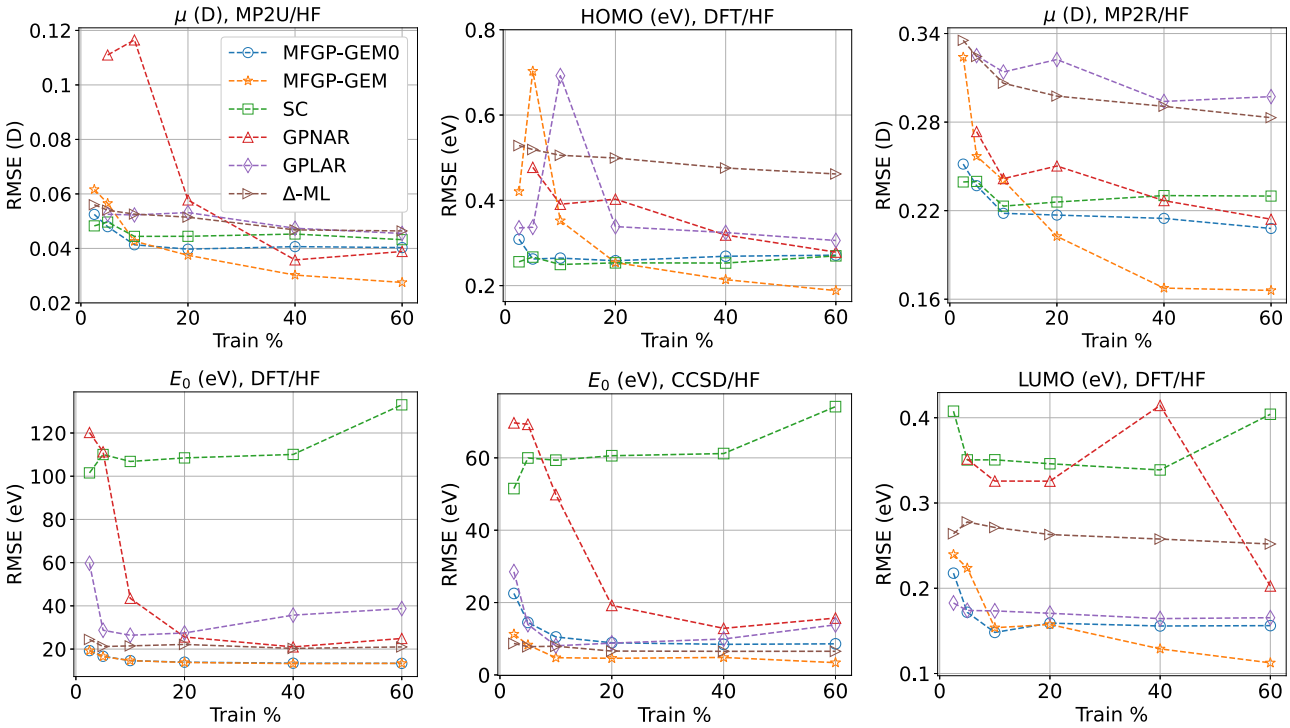

Fig. 3 | RMSE values relating to selected quantities in the Quinone data set, with the combinations of lo-fi:hi-fi indicated.

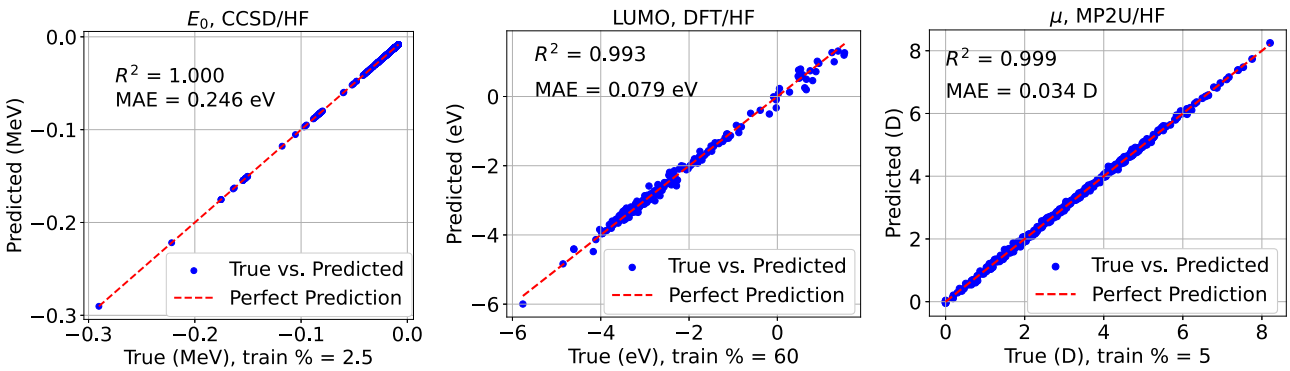

Fig. 4 | Scatterplots of true *vs.* predicted using MFGP-GEM in three cases related to the Quinone data set.

MFGP-GEM yields the lowest RMSE in all cases when at least 15–20% of the data is used for training; 20% is equivalent to 172 points for E_0, HOMO and LUMO, and 112 points for μ. Moreover, MFGP-GEM0 and MFGP-GEM are more consistent than the other methods, with MFGP-GEM being the most accurate except for very low training ratios. From this point onwards we do not show the results for MFGP-GEM0, which was not as accurate as MFGP-GEM unless the training number was very low. To visualize the levels of fit, we refer to Fig. 4, which shows scatterplots of true vs. predicted values for three cases. More cases are shown in Supplementary Fig. 4 and comparisons to other methods are shown in Supplementary Fig. 5. The RMSE values in the case of E_0 are remarkably low; ca. 5 eV for CCSD:HF, 8 eV for MP2:HF and 14 eV for DFT:HF at a 20% training ratio, which are 0.02%, 0.03% and 0.06% of the mean CCSD values. The RMSE values for μ are ca. 0.16 D for DFT:HF, 0.04 D for MP2U:HF and 0.20 D for MP2R:HF, which are 4.8%, 1.2% and 5.9% of the mean MP2R values. HOMO and LUMO require higher numbers of training points (DFT:HF) to reach similar accuracies. The RMSE values at a 60% training ratio are ca. 5.5% and 2.8 % of the mean DFT value, respectively.

HOMO and LUMO are especially challenging to calculate accurately because they involve specific orbitals. The orbital energies in HF are directly related to the eigenvalues of the Fock operator, which represents the kinetic and electrostatic interactions of the electrons, while in DFT the orbital energies are derived from the KS eigenvalues, namely the orbital energies associated with non-interacting KS orbitals. Thus, HF neglects electron correlations beyond a mean field level, while DFT incorporates electron correlation effects via the XC functional. For some materials, therefore, the different treatments may lead to a weaker correlation between quantities such as HOMO and LUMO derived from HF/post-HF methods and DFT. The same difficulty in capturing HOMO and LUMO was encountered on the QM7b data set, for which the results are shown in Supplementary Section 9.2, with MFGP-GEM again exhibiting the lowest errors. In the latter results, there appears to be a somewhat weak correlation between the DFT and GW data, probably due to the different treatments of orbital energies.

The values of μ from HF are likely to be better correlated with MP2U values than they are with the MP2R values. In MP2R, the HF geometry is relaxed, which can lead to changes in bond lengths and angles, including dihedral angles. μ depends on the overall distribution of charge within the molecule, influenced by factors such as the atom electronegativities, the molecule geometry, and the presence of polar functional groups. *Ab*-initio

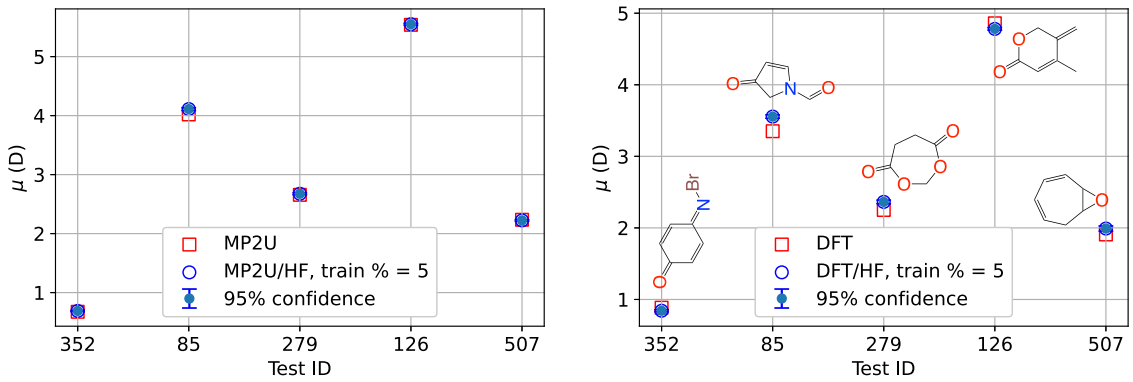

Fig. 5 | Predictions of the dipole moment μ on the Quinone data set for 5 randomly selected test points: 4-bromoiminocyclohexa-2,5-dien-1-one C_6H_4BrNO (352); 3-oxo-2H-pyrrole-1-carbaldehyde $C_5H_5NO_2$ (85); 1,3-dioxepane-4,7-dione $C_5H_6O_4$ (279); 4-methyl-5-methylidenepyran-2-one $C_7H_8O_2$ (126); 8-oxabicyclo[5.1.0] octa-2,4-diene C_7H_8O (507).

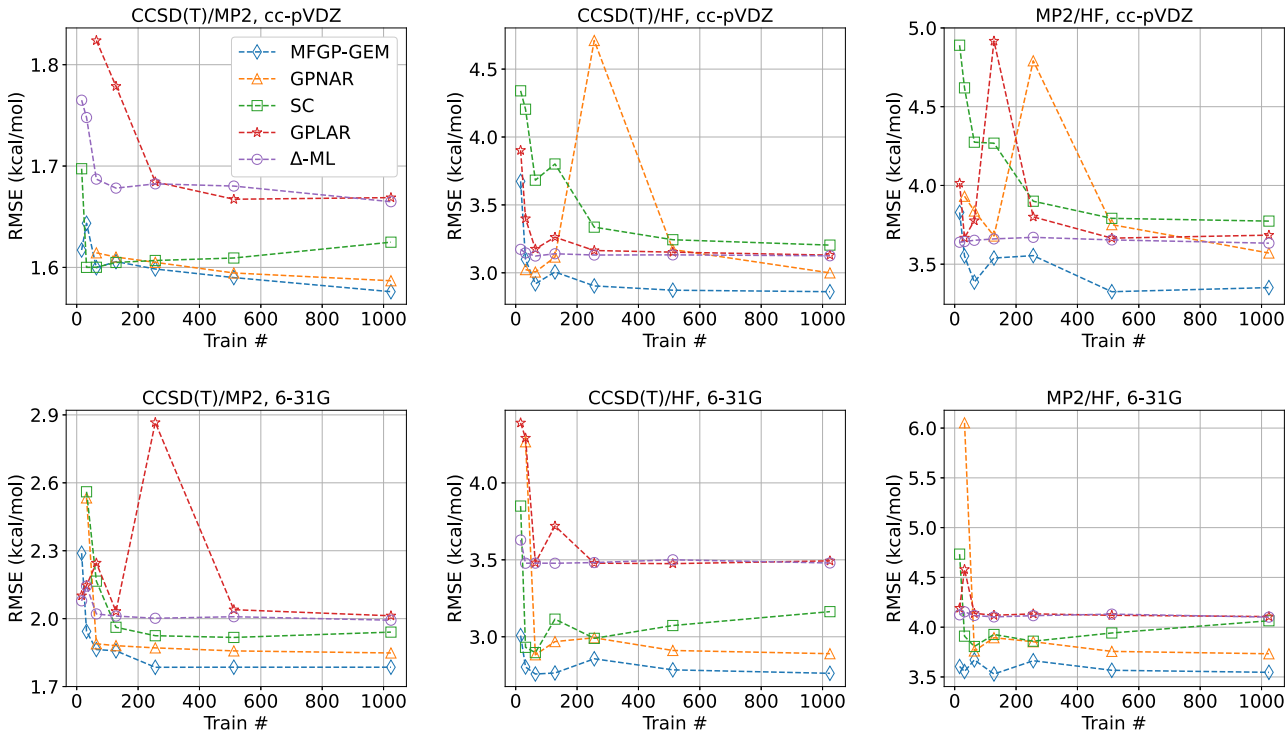

Fig. 6 | RMSE values pertaining to the prediction of E_{eff} in the QM7bE data set, with the combinations of lo-fi:hi-fi data and the basis set indicated.

approximations of μ in general differ from experimental values (X-ray diffraction) due to quantum nuclear effects, although high levels of theory such as CCSD(T) can provide reliable estimates in some cases.

Examples predictions of μ on 5 test examples selected at random when using just 5% of the data for training (43 points) are shown in Fig. 5. The 95% confidence intervals are obtained from the predictive variance $\sigma^2_{\text{predict}}$ using $y_{\text{predict}} \pm 1.96\sigma_{\text{predict}}$. Based on the additivity and transferability properties of the dipole moment, we can make some general comments on these values of μ. The presence of the highly polar carbonyl C=O groups in 4-methyl-5-methylidenepyran-2-one and 1,3-dioxepane-4,7-dione leads to relatively large values. The Br atom in 4-Bromoiminocyclohexa-2,5-dien-1-one is not as electronegative as O or N, and the imino group (NH) is not as polar as the carbonyl group, leading to a lower value of μ. Despite the relatively broad spread of μ values, the predictions are very accurate, especially in the case of MP2U:HF for the reasons outlined above.

QM7bE data set. The RMSE values relating to predictions of E_{eff} in the QM7bE data set are shown in Fig. 6. As with the Quinone and QM7b data sets, we observe that MFGP-GEM provides the most consistent and accurate results; here PCA-MACCS with 50 principal components was used as the additional descriptor. Corresponding scatterplots are shown Fig. 7. In Supplementary Figs. 9 and 10 we show the equivalent RMSE values and corresponding scatterplots for the STO-3G basis, while comparisons to other methods are shown in Supplementary Fig. 11.

The results with MFGP-GEM are highly accurate for all bases, especially for the 6-31G basis, with a minimum R^2 of 0.94 across all combinations of lo-fi:hi-fi data, basis set and training point number. The map from MP2 to CCSD(T) is the most accurate, with a minimum R^2 of 0.985. What is particularly noticeable from Fig. 6 and Supplementary Fig. 9 is that only between 16 to 128 lo-fi and hi-fi data points are required for such accuracies, e.g., to reach MAEs of 1.29 kcal/mol and 1.48 kcal/mol (within chemical accuracy) for CCSD(T):MP2(cc-pVDZ) and CCSD(T):MP2(6-31G) requires only 16:16 and 32:32 training points, respectively. For CCSD(T):HF(cc-pVDZ) and CCSD(T):HF(6-31G), MAE values of 2.3 kcal/mol and 2.23 kcal/mol were obtained using 64:64 and 32:32 training points, respectively.

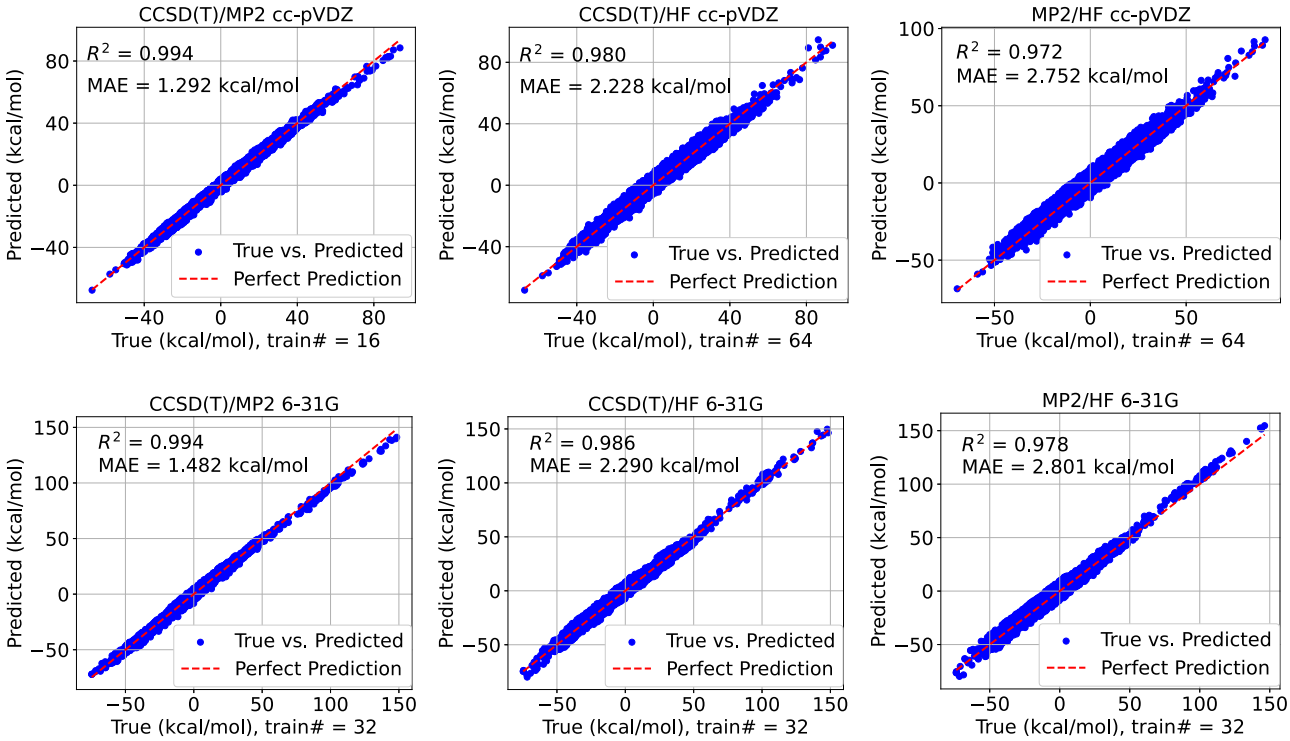

Fig. 7 | Scatterplots of the true *vs.* predicted E_{eff} using MFGP-GEM on the QM7bE data set, with different combinations of lo-fi:hi-fi data and different basis sets.

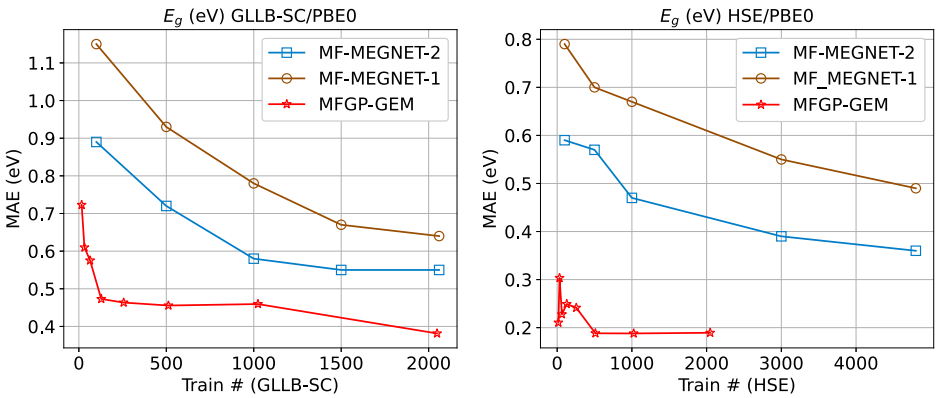

Fig. 8 | MAE values relating to predictions on the Bandgap data set with a comparison to MF-MEGNET. The results are shown for two cases: with GLLB-SC as the hi-fi and with HSE as the hi-fi, using DFT-PBE0 as the lo-fi in both cases. The number of hi-fi points is varied along the horizontal axis. MF-MEGNET-1 and MF-MEGNET-2 refer to 5000 and 41,000 lo-fi PBE0 data points used for training, respectively. For MFGP-GEM the numbers of lo-fi and hi-fi points are equal.

ML-MEGNET and CQML again failed to provide reasonable accuracies using 1:1 low-fi:hi-fi training point ratios. Instead, we used the values from the original papers at 2:1 and 4:1 ratios[54]. Report an MAE value of ~1 kcal/mol with 1024:2048 or 512:2048 training points using ratios of 1:2 or 1:4 for CCSD(T):MP2(cc-pVDZ), respectively. For CCSD(T):HF(cc-pVDZ), the same level of accuracy required 512:2048 or 128:2048 training points[55]. Found that 1024:2048 or 512:2048 CCSD(T):MP2(6-31G) training points were required to reach this level of accuracy. For the larger fidelity gap CCSD(T):HF(6-31G), their method required 1024:4096 or 512:8192 points.

Bandgap and Alexandria data sets

The MAE values on the Bandgap data set are shown in Fig. 8. (ref. 54 did not provide RMSE values in their work). Δ-ML and CQML were not implemented since there is no obvious way to generate a suitable descriptor for these methods using the available data. As far as we are aware, MF-MEGNET attains the lowest error on the Bandgap data set in the literature. Fidelities for the Bandgap data are defined by the functional, namely GLLB-SC or HSE for hi-fi and PBE0 for lo-fi. MFGP-GEM is far superior in both cases, especially considering that the best case in Fig. 8 for MF-MEGNET requires 41k PBE0 (lo-fi) points. For HSE:PBE0, MFGP-GEM attains an MAE of 0.18 eV at 512:512 points, and the lowest value attained by MF-MEGNET is 0.36 eV at 41k:4800 points. The comparison for GLLB-SC:PBE0 is similar.

Properties such as specific heat capacity rely on the distribution of electronic energy levels and vibrational frequencies, which can be estimated with relatively high accuracy using advanced composite methods. The MFGP-GEM RMSE values on the Alexandria data set are shown in Fig. 9 and Supplementary Fig. 12 for various combinations of lo-fi:hi-fi data, together with results for Δ-ML. MF-MEGNET and CQML again failed. For example, MF-MEGNET (CQML) yielded RMSE values of 103 (1277) J/mol/K and 4.05 (13.1) D for C_V and μ using all of the lo-fi data and 1024 hi-fi points in the case of G2:CBS-QB3. We note that the DFT results used a B3YLP functional and an aug-cc-pVTZ basis, while HF used a 6-311G** basis. 'Exp' denotes experimental data in Supplementary Fig. 12. At 64:64 (1024:1024) G4:CBS-QB3 training points, MFGP-GEM yields RMSE and

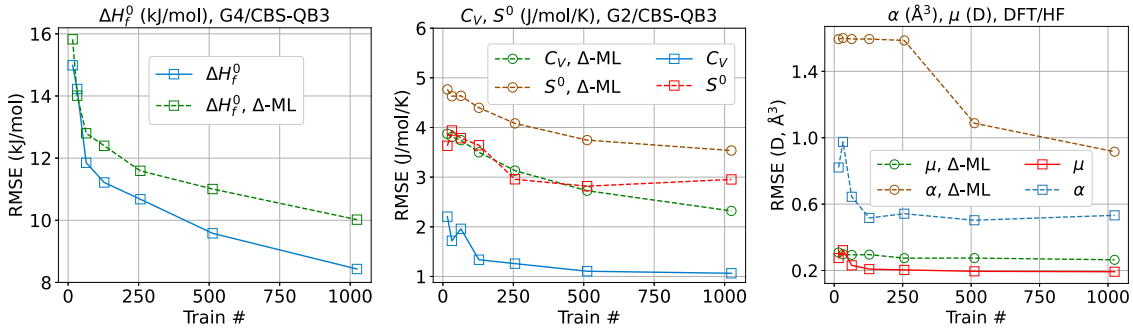

Fig. 9 | RMSE values relating to predictions in selected cases from the Alexandria data set, with the combinations of lo-fi and hi-fi data indicated.

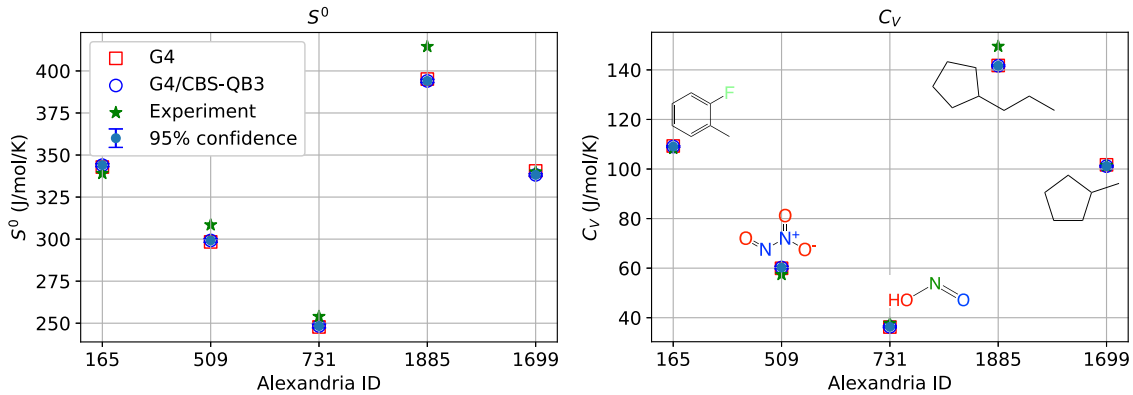

Fig. 10 | Predictions of C_V and S^0 from the Alexandria data set for 5 randomly selected test points: 1-fluoro-2-methylbenzene C_7H_7F (165), N-oxonitramide N_2O_3 (509), nitrous acid HNO_2 (731), propylcyclopentane C_8H_{16} (1885); methylcyclopentane C_6H_{12} (1699).

MAE values of 11.8 (8.4) and 8.1 (4.9) kJ/mol/K, the latter of which is 3.2% (1.9%) of the G4 mean absolute value. Supplementary Fig. 13 shows the corresponding scatterplot for only 16:16 training points, for which the R^2 is 0.997. There are several outliers in the data with values above 1000 kJ/mol/K that are nevertheless extremely well approximated, in particular the value at around 5400 kJ/mol/K. In the case of W1BD:CBS-QB, the accuracy of the ΔH_f^0 predictions is even higher. The MAE for 64:64 (512:512) training points is 5.2 (3.2) kJ/mol/K, representing 1.8% (1.1%) of the mean absolute value. Supplementary Fig. 13 shows that outliers are again extremely well approximated with only 16:16 points.

The best approximation is achieved for C_V, with all three combinations of hi-fi:lo-fi. In the case of G4:CBS-QB3, the R^2 is 1.000 and the MAE is 0.45 J/mol/K (0.13% of the mean) at 16:16 training points. Only slightly less accurate are the predictions of S^0, with MAE values of 0.78, 0.80 and 1.93 J/mol/K at 16:16 training points for W1BD:CBS-QB3, G4:CBS-QB3 and G2:CBS-QB3 (0.27%, 0.23% and 0.58% of the mean absolute value for the respective hi-fi). The results for μ are the least accurate, for the reasons discussed earlier. For DFT:HF and Exp:DFT, the MAEs at 128:128 points are 0.13 and 0.26 D, which are 7.3% and 13.4% of the respective hi-fi mean values. The approximations of α are much more accurate, with equivalent MAE values of 0.30 and 0.24 Å^3, which are 2.6% and 1.9% of the mean values.

Predictions of C_V and S^0 and the confidence intervals for 5 randomly selected test points are shown in Fig 10, alongside the experimentally measured values. Equivalent predictions for α and μ are provided in Supplementary Fig. 14. C_V and S^0 values are influenced by the size of the molecule and the number of degrees of freedom. Popylcyclopentane and methylcyclopentane are aliphatic compounds with flexible acyclic structures. The numerous degrees of freedom associated with rotations and vibrations in aliphatic molecules lead to increased heat capacities and molar entropies. In contrast, aromatic compounds such as 1-fluoro-2-methylbenzene generally have lower C_V and S^0 values due to the more constrained molecular motion. Naphthalene-1-carbaldehyde has a relatively high polarizability due to the substantial number of π electrons in the highly-conjugated aromatic naphthalene ring. The high polarizability of 1-bromooctane is primarily attributable to the presence of the halogen bromine atom and the long carbon chain. Even with this wide variety of molecules, the predictions are accurate, especially for the thermodynamic properties.

N − fidelity simulations

Both[54] and[55] found that adding more fidelities boosted the accuracy of their multi-fidelity method. We also found this to be the case in most of the examples considered. An illustration of the improvement is provided in Fig. 11, which examines cases with up to 5 fidelities (labeled *N*Fi, $N = 2, \ldots, 5$). For the definitions of cases 1 and 2 we refer to the caption of Fig. 11. In relation to the QM7bE data, the MAE (R^2) reaches 1.12 kcal/mol (0.996) and 1.04 kcal/mol (0.997) for the 6-31G and cc-pVDZ bases, respectively, using 64:64:64 CCSD(T):MP2:HF training points. In the case of STO-3G it reaches 2.21 kcal/mol, which is still very accurate. Scatterplots for the cc-pVDZ and 6-31G cases are shown in Supplementary Fig. 15. The 4Fi models lead to even higher accuracies, with MAEs of 0.99 kcal/mol for both case 1 and case 2 when using 64 points at each fidelity, and 0.91 kcal/mol when using 512 points at each fidelity. The 5Fi models yield even more dramatic improvements, with MAE values of 0.72 kcal/mol in both cases when using 64 points at each fidelity and 0.61 kcal/mol when using 512 points at each fidelity. The dipole moment prediction on the Quinone data set is also significantly improved, with a reduction in the MAE from 0.217 D at DFT:MPR to 0.137 D at HF:DFT:MPR using 20% of the data for training. At 10%, the E_0 RMSE is reduced from 0.33 eV using CCSD:DFT to 0.09 eV using HF:DFT:CCSD. As with adding more training points of the same type, adding further fidelities will eventually lead to diminishing returns.

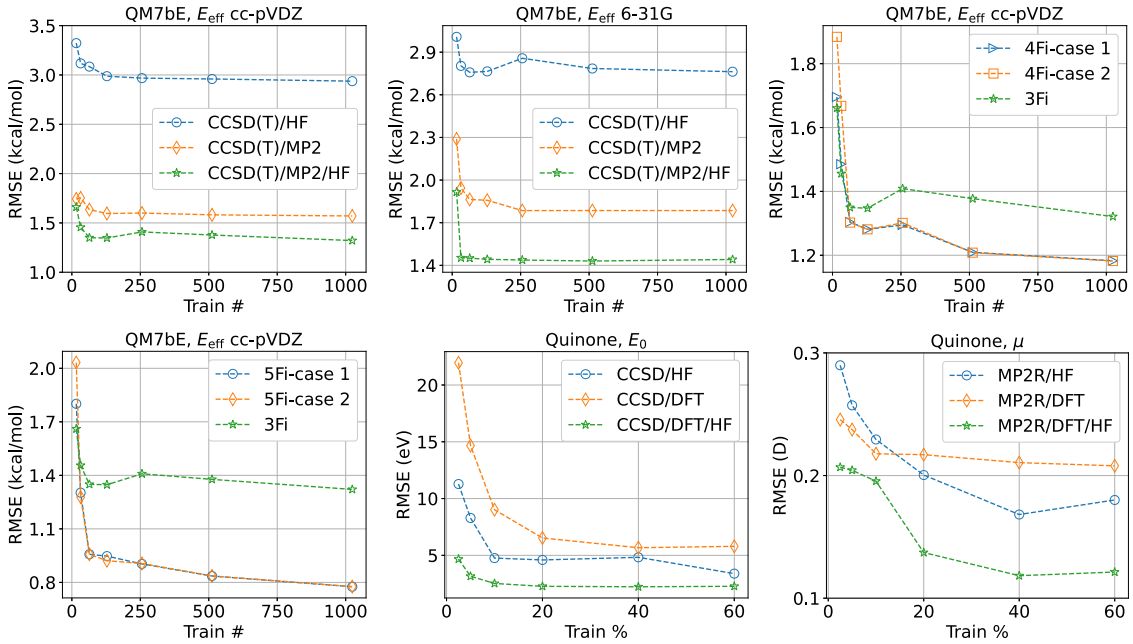

Fig. 11 | RMSE values for six different cases when using more than 2 fidelities. The quantity and fidelities are indicated in the legends and titles. 4Fi case 1 refers to the fidelities HF-STO-3G, HF-cc-pVDZ, MP2-cc-pVDZ and CCSD(T)-cc-pVDZ. 4Fi case 2 refers to HF-STO-3G, HF-6-31G, MP2-6-31G and CCSD(T)-cc-pVDZ. 5Fi case 1 refers to HF-cc-pVDZ, MP2-6-31G. MP2-cc-pVDZ, CCSD(T)-6-31G and CCSD(T)-cc-pVDZ. 5Fi case 2 refers to HF-STO-3G, HF-cc-pVDZ, MP2-STO-3G. MP2-cc-pVDZ and CCSD(T)-cc-pVDZ.

Discussion

High-fidelity QM simulation methods are generally not suitable for the rapid screening or design of materials. At the very least, such applications would require massive computational resources and budgets. Even when such resources are available, the simulation times could easily be prohibitive for complex molecules, large numbers of molecules, and costly hi-fi methods. Direct ML approaches can suffer from the same problem, since they usually require a large volume of (possibly hi-fi) data. MF methods could well provide a solution in such cases, provided that the lo-fi results are cheap to obtain, which is certainly the case with semi-empirical, HF and some DFT approaches.

In this paper, we leveraged the power of spectral embeddings and autoregressive (Markov) approaches to develop a MF method that is capable of accurate predictions using small numbers of both lo-fi and hi-fi data. We examined its performance on a range of data sets to show that it is generalizable across fidelity combinations, types of molecules and types of output. Unsurprisingly, we find that localized properties of molecules such as μ and α can be more difficult to capture than global properties. The results also suggest that the application of MF methods to the predictions of such properties requires a more careful planning of data acquisition, in that the data from different fidelities must exhibit strong correlations. Pairing methods with shared underpinnings, such as HF and post HF, will likely lead to stronger correlations than pairing methods such as GW and DFT. The $N-$ Fidelity results show that adding more fidelities can dramatically improve the results. This does not mean, however, that it is always prudent to add more fidelities. The improved accuracy has to be weighed against the increased computational costs of acquiring the additional data. In cases where data is available at HF and say G4MP2, it may well be worth generating additional data at the DFT level. There may also be cases in which results at intermediate levels of theory can be acquired efficiently, for example in composite methods.

We investigated relatively simple adjacency matrices in this work and it is possible to use more information related to the molecules. Although we do not expect the results to improve dramatically, further accuracy gains are possible. Another route to improved accuracy is transfer learning, which would allow for different ratios of lo-fi and hi-fi training data, and therefore a fuller exploitation of the short simulations times associated with the former.

Methods

Gaussian processes based on spectral graph embeddings via diffusion maps

Backgrounds on the ML and MF methods used for comparison are provided in Supplementary Sections 5 and 6, respectively. Implementations of these methods are described in "Implementations of direct machine learning and other multi-fidelity methods".

A graph $\mathcal{G}$ is an ordered pair (V, E), where V is a non-empty set of vertices or nodes, and E is a set of edges, each of which is an unordered pair of distinct vertices from V. In the context of molecules, the vertices $v_1, \ldots, v_n \in V$ in $\mathcal{G}$ are identified with the n atoms. The atom properties, such as atom type, position and charge are used as node features. A set of edges $E = \{(v_i, v_j) | v_i, v_j \in V, v_i \neq v_j\}$ is defined to represent the bonds, indicating whether or not v_i and v_j are connected. In a weighted graph $\mathcal{G} = (V, E, W)$, each edge (v_i, v_j) is associated with a weight $w_{ij} \in W$, which represents the strength of the relationship between the associated vertices. In molecular graphs, weights (or edge features) can be assigned to the edges between atoms, based on bond properties such as the type, strength and length. Molecular graphs are undirected since the order of the vertices in the edges (v_i, v_j) is immaterial. Moreover, they are connected, since there is a path between any two distinct vertices.

Diffusion maps[61] (see Supplementary Section 4.1) define embeddings on graphs using the concept of a diffusion space. The vertices are embedded isometrically in diffusion space by preserving a diffusion distance. Let $\mathcal{G}$ be a weighted molecular graph with vertices $V = \{v_1, v_2, \ldots, v_n\}$ for a molecule m possessing n atoms. An adjacency or connectivity matrix $\mathbf{K} = [K_{ij}]_{i,j} \in \mathbb{R}^{n \times n}$ can be constructed from the weights w_{ij}, considering the presence of bonds, bond types, bond strengths or other properties (edge features) by using, e.g., thresholding for a particular feature or a weighted average of the features. The adjacency matrix entry $K_{ij} = w_{ij}$ defines a metric (distance or similarity measure) $d(v_i, v_j)$ between nodes i and j. A diffusion process on $\mathcal{G}$ is obtained from the diffusion matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{K}$, in which $\mathbf{D}$ is the degree

matrix. The matrix $\mathbf{P}$ is a Markov matrix for a random walk on the graph. t-step transition probabilities are obtained from $\mathbf{P}^t$.

The eigenvalues γ_i of $\mathbf{P}$ can be ordered such that $1 = \gamma_1 > \cdots > \gamma_n$, with corresponding right and left eigenvectors $\mathbf{r}_i \in \mathbb{R}^n$ and $\mathbf{l}_i \in \mathbb{R}^n$, respectively. The key quantities required for defining the diffusion maps are the row vectors $\mathbf{p}_j^t = \sum_{i=1}^n (\gamma_i)^t r_{ji} \mathbf{l}_i$ of $\mathbf{P}^t$, $j = 1, \ldots, n$. Here, r_{ji} is the j-th component of $\mathbf{r}_i$. The vector $\mathbf{p}_j^t$ defines a probability mass function, with the i-th component being the probability of reaching vertex v_i from v_j in t steps. Diffusion maps $\psi^t : V \to \mathcal{D}^{(t)} \subset \mathbb{R}^n$ are defined by $\psi^t(v_i) = \left((\gamma_1)^t r_{i1}, \ldots, (\gamma_m)^t r_{in}\right)^T$, mapping to a diffusion space $\mathcal{D}^{(t)}$, with preservation of a diffusion distance (Supplementary Section 4.1). The decay in the eigenvalues leads to low-dimensional embeddings $\psi_r^t(v_i) : V \to \mathcal{D}_r^{(t)} \subset \mathbb{R}^r$ by restricting $\psi^t(v_i)$ to the first $r < n$ (user chosen) components.

Graphs $\mathcal{G}_m$ can be formed for each molecule m to extract the right eigenvectors and eigenvalues, $\{\mathbf{r}_{i,m}\}_{i=1}^r$ and $\{\gamma_{i,m}\}_{i=1}^r$, respectively. Since the molecules have different numbers of atoms n, the $\mathbf{r}_{i,m}$ are padded with 0s such that they all lie in $\mathbb{R}^{\hat{n}}$, where $\hat{n}$ is the number of atoms in the largest molecule. The set of vectors and eigenvalues can then be concatenated to produce a feature vector representing molecule m

$$\mathbf{z}_m = \left(\bigoplus_{i=1}^r \{\mathbf{r}_{i,m}\}\right) \oplus \left(\bigoplus_{i=1}^r \{\gamma_{i,m}\}\right) \in \mathbb{R}^{r(\hat{n}+1)}, \quad m = 1, \ldots, M \tag{1}$$

in which $\oplus$ is used to denote a concatenation. A new graph $\widehat{\mathcal{G}} = (\widehat{V}, \widehat{E}, \widehat{W})$ is now defined by associating each node $\hat{v}_m \in \widehat{V}$ with a molecule m, for which we have a feature vector $\mathbf{z}_m$. An adjacency matrix $\widehat{\mathbf{K}} \in \mathbb{R}^{M \times M}$ can be defined using a kernel function $k(\mathbf{z}_l, \mathbf{z}_m)$, e.g., the following Gaussian kernel

$$\widehat{d}(\hat{v}_l, \hat{v}_m) = k(\mathbf{z}_l, \mathbf{z}_m) = \exp\left(-\frac{\|\mathbf{z}_l - \mathbf{z}_m\|^2}{s^2}\right) \tag{2}$$

for some shape parameter s (user chosen). This specifies the metric $\widehat{d}(\cdot,\cdot)$ and also the weights $\hat{w}_{mn} \in \widehat{W}$ associated with the edges in $\widehat{E}$. A diffusion map embedding is performed as above to yield right eigenvectors $\hat{\mathbf{r}}_m$ and eigenvalues $\hat{\gamma}_m$, alongside low-dimensional embeddings $\phi_q^t(u_j) : \widehat{V} \to \widehat{\mathcal{D}}_q^{(t)} \subset \mathbb{R}^q$ given by

$$\mathbf{x}_m = \phi_q^t(\hat{v}_m) = \left(\hat{\gamma}_1^t \hat{r}_{m1}, \ldots, \hat{\gamma}_q^t \hat{r}_{mq}\right)^T, \quad \forall m \in \{1, \ldots, M\} \tag{3}$$

for some $q < M$. $\widehat{\mathcal{D}}_q^{(t)}$ is a diffusion space and $\hat{r}_{ji}$ is the j-th component of $\hat{\mathbf{r}}_i$. Optionally, a descriptor or concatenation of descriptors $\mathbf{d}_m \in \mathbb{R}^d$ can be combined with $\mathbf{x}_m$, to produce features $\mathbf{x}_m \mapsto \mathbf{x}_m + \mathbf{d}_m \in \mathbb{R}^{q+d}$, which are fed to the multi-step nonlinear autoregressive model described below. The procedure to extract the $\mathbf{x}_m$ is illustrated in Fig. 2.

We assume the autoregressive form $y^F = \eta^F(\eta^{f \in \mathcal{F}}(\mathbf{x}), \mathbf{x}) + \epsilon(\mathbf{x})$, in which $\mathcal{F} \subseteq \{F-1, \ldots, 1\}$ is a subset of the fidelity indices. $\epsilon|\sigma \sim \mathcal{GP}\left(0, \sigma^2 \delta(\mathbf{x}, \mathbf{x}')\right)$ is the error (σ^2 is the noise variance) and $\delta(\mathbf{x}, \mathbf{x}')$ is the Kronecker delta function. We note that unlike GPNAR, this is a generalized Markov model with $\dim(\mathcal{F})$ steps; that is, $\eta^F \perp \{\eta^f : f \notin \mathcal{F}\} | \{\eta^f : f \in \mathcal{F}\}$. A GP prior $\eta^F|\theta \sim \mathcal{GP}(0, c(\cdot,\cdot|\theta))$ with covariance function $c\left([\mathbf{x}, \eta^{f \in \mathcal{F}}(\mathbf{x})], [\mathbf{x}', \eta^{f \in \mathcal{F}}(\mathbf{x}')] | \theta\right)$ containing hyperparameters θ is placed over the latent function. The notation $[\eta^{f \in \mathcal{F}}(\mathbf{x}), \mathbf{x}]$ denotes a concatenation of selected fidelities and $\mathbf{x}$. Since the latent function values $\eta^f(\mathbf{x}_m^f)$ are not known (unless the computations are error free), we approximate these values by the observations y^f and absorb the error resulting in this approximation into a redefined error. We first define a compact notation

$$\mathbf{s} = \bigoplus_{f \in \mathcal{F}} y^f \oplus \mathbf{x} = \bigoplus_{f \in \mathcal{F}} (\eta^f(\mathbf{x}) + \epsilon^f(\mathbf{x})) \oplus \mathbf{x} \tag{4}$$

The prior then takes the form $\eta^F|\theta \sim \mathcal{GP}(0, c(\mathbf{s}, \mathbf{s}'|\theta))$ for a model $y^F = \eta^F(\mathbf{s}) + \epsilon(\mathbf{s})$, in which $\mathbf{s}' = \bigoplus_{f \in \mathcal{F}} (\eta^f(\mathbf{x}') + \epsilon^f(\mathbf{x}')) \oplus \mathbf{x}'$ and $\epsilon(\mathbf{s})$ is Gaussian white noise with variance σ^2. The posterior predictive distribution is

$$\begin{aligned}
&\hat{\eta}^F(\mathbf{s}) | \{\mathbf{y}^f\}, \{\mathbf{X}^f\}, \boldsymbol{\theta} \sim \mathcal{GP}(m'(\mathbf{s}|\boldsymbol{\theta}), c'(\mathbf{s}, \mathbf{s}'|\boldsymbol{\theta})) \\
&m'(\mathbf{s}|\boldsymbol{\theta}) = \mathbf{c}(\mathbf{s}|\boldsymbol{\theta})^T (\mathbf{C}(\boldsymbol{\theta}) + \sigma^2 \mathbf{I})^{-1} \mathbf{y}^F, \\
&c'(\mathbf{s}, \mathbf{s}'|\boldsymbol{\theta}) = c(\mathbf{s}, \mathbf{s}'|\boldsymbol{\theta}) + \mathbf{c}(\mathbf{s})^T (\mathbf{C}(\boldsymbol{\theta}) + \sigma^2 \mathbf{I})^{-1} \mathbf{c}(\mathbf{s}') \\
&\mathbf{c}(\mathbf{s}) = (c(\mathbf{s}, \mathbf{s}_1|\boldsymbol{\theta}), \ldots, c(\mathbf{s}, \mathbf{s}_N|\boldsymbol{\theta}))^T
\end{aligned} \tag{5}$$

in which $\mathbf{C}(\theta) = [c(\mathbf{s}_m, \mathbf{s}_{m'}|\theta)]_{m,m'} \in \mathbb{R}^{N \times N}$ is the covariance matrix, N is the number of training points, and $\mathbf{s}_m = \bigoplus_{f \in \mathcal{F}} y_m^f \oplus \mathbf{x}_m^F$, recalling that $\mathbf{X}^f \subseteq \mathbf{X}^{f-1}$. The predictive mean and variance at a given $\mathbf{s}$ are $m'(\mathbf{s}|\theta)$ and $c'(\mathbf{s}, \mathbf{s}|\theta)$, respectively. The hyperparameters are estimated using a maximum log likelihood

$$\theta, \sigma = \text{argmax}_{\theta,\sigma} = \left(-\frac{1}{2}\ln|(\mathbf{C}(\theta) + \sigma^2\mathbf{I})| - \frac{1}{2}\mathbf{y}^T(\mathbf{C}(\theta) + \sigma^2\mathbf{I})^{-1}\mathbf{y}^F\right) \tag{6}$$

Implementation details

MFGP-GEM was implemented in Python 3.8.17 using the `GPy` package, and used a maximum likelihood to infer the hyperparameter values based on a square (L_2) loss function and a non-stationary sum of Matérn 5/2 automatic relevance determination (ARD), linear ARD and Gaussian noise kernels. We found that this kernel, or a kernel that replaced the Matérn 5/2 ARD component with a squared-exponential ARD (SEARD) kernel worked better than SEARD or Matérn ARD kernels alone, probably due to the fact that MF problems are essentially sequence problems and therefore not stationary. Structures were created with the `MolFromSmiles()` function from the `Chem` library in RDkit. From these structures, two types of adjacency matrices $\mathbf{K}$ were considered, 'bond' and 'connectivity'. 'connectivity' refers to simple binary-valued entries that indicate the presence or absence of a bond. 'bond' included both the presence and type of bond, e.g., single (1), double (2) or triple (3), with 0 used for the absence of a bond. More elaborate adjacency matrices $\mathbf{K}$ such as entries that are linear combinations of bond type encodings and atomic weights of the species in the edge could also be used.

For the QM7b data set, the bond information for the adjacency matrix was derived from the Coulomb matrices. For the QM7bE and Bandgap data sets we used the connectivity-based adjacency matrix since only `xyz` data is available. A bond distance threshold was defined to decide connectivity; those atoms separated by a distance below the threshold were considered connected. For the Alexandria data set, the available InChi representations were converted to SMILEs using the `Chem.MolFromInchi()` and `Chem.MolToSmiles()` functions in RDkit. The additional descriptors were generated in Python from SMILEs representations using `AllChem` in RDkit. All fingerprints and the SoB were assumed to have a size of 1024 bits, while MACCS keys have a size of 166 bits. To generate PCA and SPCA representations we used the scikit-learn package. ACSF and SOAP were implemented with the `DScribe` package. Where available, the atomic coordinates were used to define the centers in both methods. For each data set, the radial cutoff, number of spherical harmonics and number of radial basis functions were varied in SOAP to achieve the best results. It was found that spherical Gaussian type orbitals were superior to a polynomial basis. Similarly, for ACSF, the radial cutoff and symmetry function parameters were varied to obtain the best fits. These local descriptors were transformed into global descriptors using the averaging option. Zero padding was used to create descriptors of a fixed size, determined by the size of the largest molecule.

Generation of the Quinone data set

Initial geometries were generated from SMILEs representations using `RDkit`. All calculations were performed using the Python package PySCF[83] based on optimized geometries using unrestricted HF (UHF). Subsequent HF calculations were performed with a cc-pVDZ basis. The

Table 2 | Architectures of the GCN, MLP and CNN used for direct machine learning comparisons

Method	Architecture
GCN	GCN(in,64)-ReLU-GCN(64,64)-ReLU-FC(128)-ReLU-FC(64)-ReLU-FC(1)
CNN	Im(H,W)-conv2d([2,2],32,p)-BN-ReLU-conv2d([2,2],32,p)-BN-ReLU-flatten-FC(32)-ReLU-FC(16)-ReLU-FC(1)
MLP	FC(in,128)-ReLU-FC(64)-ReLU-FC(32)-ReLU-FC(1)

Notation is as follows: 'GCN(a,b)' is a GCN layer with input size 'a' and hidden space dimension 'b'; 'in' is the size of the input space; 'FC(c)' is a fully-connected (dense) layer with 'c' nodes; 'ReLU' is the action of a Rectified Linear Unit function; 'Im(H,W)' is an image input layer with height H and width W; 'conv2d([2,2],32,p)' is a 2-d convolutional layer with 32 filters each of size [2,2] and with padding to preserve the size; 'BN' is a batch normalization layer; 'flatten' is a flattening operation; 'FC(in,128)' is a dense layer with 128 nodes and an input space size of 'in'.

HOMO energy in HF is determined from $E^{\mathrm{HF}}_{\mathrm{HOMO}} = \varepsilon^{\mathrm{HF}}_{\mathrm{occupied}}$, and the LUMO from $E^{\mathrm{HF}}_{\mathrm{LUMO}} = \varepsilon^{\mathrm{HF}}_{\mathrm{unoccupied}}$. DFT calculations were performed with unrestricted KS using the B3LYP XC functional. HOMO and LUMO indices were identified from the KS orbital energies and occupancies. MP2 and CCSD calculations used the UHF results as reference wavefunctions. Density-Fitted MP2 was used to obtain the unrelaxed and relaxed dipole moments. Full details of the generation of the data set are provided in Supplementary Section 2.1.

Implementations of direct machine learning and other multi-fidelity methods

The direct GP model, MEGNET and the GCN were implemented in Python, while the CNN and MLP models were implemented in Matlab 2023a. The GP model used GPy, with a maximum likelihood for inferring hyperparameters. The mean values were assumed to be zero, and a SEARD kernel with Gaussian noise was employed. All networks used the square error (L_2) loss function and were trained with the Adam optimizer[84]. The GCN, MLP and CNN architectures described below were found to be optimal using initial hand tuning followed by stochastic grid searches. Initial learning rates of 10^{-3} were used in all cases.

The GCN was implemented in PyTorch using the `GCNConv` function in PyTorch Geometric. The architecture used is summarized in Table 2. The size of the input space 'in' differs according to the data set. For the QM7b data set, padded Coulomb matrices were used to create the graph via the `nx.Graph()` function in NetworkX. Node features for the $j-$th atom were set equal to the values in the $j-$th row of the Coulomb matrix, which describe interactions between the $j-$th atom and the other atoms in the molecule. An edge feature between nodes j and k was defined by the $(j,k)-$th value of the Coulomb matrix, which represents an interaction strength. The same procedure was used for the Quinone data set, with Coulomb matrices generated from SMILES using our own custom functions based on RDkit. For the QM7bE data set, pymatgen structures were used to generate the Coulomb matrices.

The direct MEGNET model was implemented according to the recommended structure described in Supplementary Section 5.5. We chose $r_{\mathrm{cutoff}} = 5$ and $n = 100$ for the cutoff distance in the `GaussianDistance` function and the number of features, respectively. For the Quinone data set, the MEGNET `get_pmg_mol_from_smiles` function was used to convert SMILES representations to pymatgen structures in order to use `CrystalGraph`. For the QM7bE data set, structures were created from available `xyz` files[55] using the pymatgen `XYZ` and `Molecule` functions, based on the atomic coordinates and species. The QM7b data set provides only Coulomb matrices, from which we were unable to generate structures.

The CNN architecture is summarized in Table 2. The height H and width W correspond to the number of descriptors and lengths of the descriptor sequences, respectively (we may think of the descriptors as vectors in $\mathbb{R}^d$ or equivalently as sequences with d terms). In the case of full sequences, rather than PCA versions, the MACCS keys were padded with 0's to be of length 1024 if they were used in combination with other descriptors.

The MLP architecture is also shown in Table 2. The input was a descriptor or concatenation of descriptor sequences, resulting in a length 'in'.

The GP multi-fidelity models were again implemented in Python using GPy. GPNAR and GPLAR followed their original formulations[51,57], employing the SEARD kernel with Gaussian noise. Stochastic collocation was also implemented according to the original formulation[58]. MF-MEGNET was implemented according to the original work[54] by appending a fidelity index to the structures, namely 1 for low and 2 for high, under the 'state' key. These variables are passed as global features to MEGNET.

Data availability

The Quinone data set is available on Github from https://github.com/ashah1973/MFGP-GEM/tree/main. The Bandgap data set can be downloaded from flagshare using the link https://figshare.com/articles/dataset/Learning_Properties_of_Ordered_and_Disordered_Materials_from_Multi-fidelity_Data/13040330 provided by[54]. The corresponding structures for the Bandgap data are available from the Materials project using the `MPRester` function. Note that some of the IDs have since been changed or the data has been removed.

Code availability

The MFGP-GEM code for implementing the embeddings, low-dimensional data representations and multi-fidelity model are available on Githib from https://github.com/ashah1973/MFGP-GEM/tree/main.

Received: 17 May 2024; Accepted: 12 November 2024;
Published online: 02 March 2025

References

1. Pham, T.-H., Qiu, Y., Zeng, J., Xie, L. & Zhang, P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to covid-19 drug repurposing. *Nat. Mach. Intell.* **3**, 247–257 (2021).
2. Zerner, M. C., Loew, G. H., Kirchner, R. F. & Mueller-Westerhoff, U. T. An intermediate neglect of differential overlap technique for spectroscopy of transition-metal complexes. ferrocene. *J. Am. Chem. Soc.* **102**, 589–599 (1980).
3. Roothaan, C. C. J. New developments in molecular orbital theory. *Rev. Mod. Phys.* **23**, 69 (1951).
4. Sherrill, C. D. & Schaefer III, H. F. The configuration interaction method: advances in highly correlated approaches, *Adv. Quantum Chem*. **34**, 143–269 (1999).
5. Møller, C. & Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **46**, 618 (1934).
6. Shavitt, I. & Bartlett, R. J. *Many-body methods in chemistry and physics: MBPT and coupled-cluster theory* (Cambridge university press, 2009).
7. Aryasetiawan, F. & Gunnarsson, O. The gw method. *Rep. Prog. Phys.* **61**, 237 (1998).
8. Parr, R. G., Gadre, S. R. & Bartolotti, L. J. Local density functional theory of atoms and molecules. *Proc. Natl Acad. Sci. USA* **76**, 2522–2526 (1979).
9. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, 864–871 (1964).
10. Parr, R. G. Density functional theory of atoms and molecules. in *Horizons of quantum chemistry* (Springer, 1980).
11. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
12. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).
13. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *J. Chem. Phys.* **127**, 124105 (2007).

14. Stewart, R. F. Small gaussian expansions of slater-type orbitals. *J. Chem. Phys.* **52**, 431–438 (1970).
15. McGrath, M. P. & Radom, L. Extension of gaussian-1 (g1) theory to bromine-containing molecules. *J. Chem. Phys.* **94**, 511–516 (1991).
16. Dunning Jr, T. H. Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen. *J. Chem. Phys.* **90**, 1007–1023 (1989).
17. Rosen, A. S. et al. High-throughput predictions of metal–organic framework electronic properties: theoretical challenges, graph neural networks, and data exploration. *npj Comput. Mater.* **8**, 112 (2022).
18. Elton, D. C., Boukouvalas, Z., Butrico, M. S., Fuge, M. D. & Chung, P. W. Applying machine learning techniques to predict the properties of energetic materials. *Sci. Rep.* **8**, 9059 (2018).
19. Reiser, P. et al. Graph neural networks for materials science and chemistry. *Commun. Mater.* **3**, 93 (2022).
20. Hansen, K. et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
21. Schmidt, J., Wang, H.-C., Schmidt, G. & Marques, M. A. Machine learning guided high-throughput search of non-oxide garnets. *npj Comput. Mater.* **9**, 63 (2023).
22. Kirkpatrick, J. et al. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science* **374**, 1385–1389 (2021).
23. Margraf, J. T. & Reuter, K. Pure non-local machine-learned density functional theory for electron correlation. *Nat. Commun.* **12**, 1–7 (2021).
24. Ellis, J. A. et al. Accelerating finite-temperature kohn-sham density functional theory with deep neural networks. *Phys. Rev. B* **104**, 035120 (2021).
25. Wang, L. et al. Quantum chemical descriptors in quantitative structure-activity relationship models and their applications. *Chemom. Intell. Lab. Syst.* **217**, 104384 (2021).
26. Manzhos, S. & Carrington Jr, T. A random-sampling high dimensional model representation neural network for building potential energy surfaces. *J. Chem. Phys.* **125**, 084109 (2006).
27. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
28. Faber, F. A. et al. Prediction errors of molecular machine learning models lower than hybrid dft error. *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).
29. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B-Condens. Matter Mater. Phys.* **87**, 184115 (2013).
30. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
31. Himanen, L. et al. Dscribe: library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
32. Jäger, M. O., Morooka, E. V., Federici Canova, F., Himanen, L. & Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Comput. Mater.* **4**, 37 (2018).
33. Lange, J. J. et al. Comparative analysis of chemical descriptors by machine learning reveals atomistic insights into solute-lipid interactions. *Mol. Pharm.* **21**, 3343–3355 (2024).
34. Santiago, R., Vela, S., Deumal, M. & Ribas-Arino, J. Unlocking the predictive power of quantum-inspired representations for intermolecular properties in machine learning. *Digital Discov.* **3**, 99–112 (2024).
35. Xu, P., Ji, X., Li, M. & Lu, W. Small data machine learning in materials science. *npj Comput. Mater.* **9**, 42 (2023).
36. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
37. Dou, B. et al. Machine learning methods for small data challenges in molecular science. *Chem. Rev.* **123**, 8736–8780 (2023).
38. Yamada, H. et al. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.* **5**, 1717–1730 (2019).
39. Kong, S., Guevarra, D., Gomes, C. P. & Gregoire, J. M. Materials representation and transfer learning for multi-property prediction. *Appl. Phys. Rev.* **8**, 021409 (2021).
40. Lee, J. & Asahi, R. Transfer learning for materials informatics using crystal graph convolutional neural network. *Comput. Mater. Sci.* **190**, 110314 (2021).
41. Hoffmann, N., Schmidt, J., Botti, S. & Marques, M. A. Transfer learning on large datasets for the accurate prediction of material properties. *Digital Discov.* **2**, 1368–1379 (2023).
42. Wu, S. et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *Npj Comput. Mater.* **5**, 66 (2019).
43. Zhang, Y. et al. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* **10**, 8154–8163 (2019).
44. Hayes, N., Merkurjev, E. & Wei, G.-W. Integrating transformer and autoencoder techniques with spectral graph algorithms for the prediction of scarcely labeled molecular data. *Comput. Biol. Med.* **153**, 106479 (2023).
45. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
46. Schütt, K., Sauceda, H., Kindermans, P., Tkatchenko, A. & Müller, K. Schnet-a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
47. Gasteiger, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs, arXiv:2003.03123 (2022).
48. Gasteiger, J., Yeshwanth, C. & Günnemann, S. Directional message passing on molecular graphs via synthetic coordinates. *Adv. Neural Inf. Process. Syst.* **34**, 15421–15433 (2021).
49. Fung, V., Zhang, J., Juarez, E. & Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *npj Comput. Mater.* **7**, 84 (2021).
50. Peherstorfer, B., Willcox, K. & Gunzburger, M. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Rev.* **60**, 550–591 (2018).
51. Kennedy, M. C. & O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **87**, 1–13 (2000).
52. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the δ-machine learning approach. *J. Chem. theory Comput.* **11**, 2087–2096 (2015).
53. Huang, B., von Lilienfeld, O. A., Krogel, J. T. & Benali, A. Toward dmc accuracy across chemical space with scalable δ-qml. *J. Chem. Theory Comput.* **19**, 1711–1721 (2023).
54. Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* **1**, 46–53 (2021).
55. Zaspel, P., Huang, B., Harbrecht, H. & von Lilienfeld, O. A. Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited. *J. Chem. Theory Comput.* **15**, 1546–1559 (2019).
56. Fare, C., Fenner, P., Benatan, M., Varsi, A. & Pyzer-Knapp, E. O. A multi-fidelity machine learning approach to high throughput materials screening. *npj Comput. Mater.* **8**, 257 (2022).
57. Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N. & Karniadakis, G. E. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **473**, 20160751 (2017).
58. Narayan, A., Gittelson, C. & Xiu, D. A stochastic collocation algorithm with multifidelity models. *SIAM J. Sci. Comput.* **36**, A495–A521 (2014).

59. Egorova, O., Hafizi, R., Woods, D. C. & Day, G. M. Multifidelity statistical machine learning for molecular crystal structure prediction. *J. Phys. Chem. A* **124**, 8065–8078 (2020).
60. Tran, A., Tranchida, J., Wildey, T. & Thompson, A. P. Multi-fidelity machine-learning with uncertainty quantification and bayesian optimization for materials design: application to ternary random alloys. *J. Chem. Phys.* **153**, 074705 (2020).
61. Donoho, D., Chui, C., Coifman, R. R. & Lafon, S. Special issue: diffusion maps and Wavelets Diffusion maps. *Appl. Comput. Harmonic Anal.* **21**, 5 – 30 (2006).
62. Coifman, R. R. et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl Acad. Sci. USA* **102**, 7426–7431 (2005).
63. Venkitaraman, A., Chatterjee, S. & Handel, P. Gaussian processes over graphs. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020).
64. Borovitskiy, V. et al. Matérn gaussian processes on graphs. In *International Conference on Artificial Intelligence and Statistics* (PMLR, 2021).
65. Gao, P. et al. Graphical gaussian process regression model for aqueous solvation free energy prediction of organic molecules in redox flow batteries. *Phys. Chem. Chem. Phys.* **23**, 24892–24904 (2021).
66. Cramer, C. J. *Essentials of computational chemistry: theories and models* (John Wiley & Sons, 2013).
67. Montavon, G. et al. Machine learning of molecular electronic properties in chemical compound space. *N. J. Phys.* **15**, 095003 (2013).
68. Tkatchenko, A., DiStasio Jr, R. A., Car, R. & Scheffler, M. Accurate and efficient method for many-body van der waals interactions. *Phys. Rev. Lett.* **108**, 236402 (2012).
69. Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The pbe0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).
70. Ghahremanpour, M. M., Van Maaren, P. J. & Van Der Spoel, D. The alexandria library, a quantum-chemical database of molecular properties for force field development. *Sci. Data* **5**, 1–10 (2018).
71. Pople, J. A., Head-Gordon, M., Fox, D. J., Raghavachari, K. & Curtiss, L. A. Gaussian-1 theory: a general procedure for prediction of molecular energies. *J. Chem. Phys.* **90**, 5622–5629 (1989).
72. Curtiss, L. A., Raghavachari, K., Trucks, G. W. & Pople, J. A. Gaussian-2 theory for molecular energies of first-and second-row compounds. *J. Chem. Phys.* **94**, 7221–7230 (1991).
73. Barnes, E. C., Petersson, G. A., Montgomery Jr, J. A., Frisch, M. J. & Martin, J. M. Unrestricted coupled cluster and brueckner doubles variations of w1 theory. *J. Chem. Theory Comput.* **5**, 2687–2693 (2009).
74. Montgomery Jr, J. A., Frisch, M. J., Ochterski, J. W. & Petersson, G. A. A complete basis set model chemistry. vi. use of density functional geometries and frequencies. *J. Chem. Phys.* **110**, 2822–2827 (1999).
75. Montgomery Jr, J. A., Frisch, M. J., Ochterski, J. W. & Petersson, G. A. A complete basis set model chemistry. vii. use of the minimum population localization method. *J. Chem. Phys.* **112**, 6532–6542 (2000).
76. Gritsenko, O., van Leeuwen, R., van Lenthe, E. & Baerends, E. J. Self-consistent approximation to the kohn-sham exchange potential. *Phys. Rev. A* **51**, 1944 (1995).
77. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).
78. 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.* **108**, 058301 (2012).
79. Durant, J. L., Leland, B. A., Henry, D. R. & Nourse, J. G. Reoptimization of mdl keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280 (2002).
80. Hansen, K. et al. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
81. Hall, L. H. & Kier, L. B. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **35**, 1039–1045 (1995).
82. Zou, H., Hastie, T. & Tibshirani, R. Sparse principal component analysis. *J. Comput. Graph. Stat.* **15**, 265–286 (2006).
83. Sun, Q. et al. Pyscf: the python-based simulations of chemistry framework. *WIREs Comput. Mol. Sci.* **8**, e1340 (2018).
84. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

Author contributions

A.A.S. conceived the project and the MFGP-GEM method, implemented the ML, MF and electronic-structure models, performed the simulations to generate the results, conducted the analysis and wrote the first draft. P.K.L. assisted on implementing the electronic structure methods, analyzing the results and editing the manuscript. W.W.X. assisted on implementing the MF and ML methods, analyzing the results and editing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01479-0.

Correspondence and requests for materials should be addressed to Akeel A. Shah or W. W. Xing.