Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d5sc09780k

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 12th December 2025
Accepted 22nd February 2026

DOI: 10.1039/d5sc09780k

rsc.li/chemical-science

Transfer learning of GW Bethe–Salpeter equation excitation energies

Dario Baum, Arno Förster and Lucas Visscher *

A persistent challenge in machine learning for electronic-structure calculations is the sharp imbalance between abundant low-fidelity data like (time-dependent) density functional theory [(TD)DFT] results and the scarcity of high-fidelity data like many-body perturbation theory labels. We show that transfer learning provides an effective route to bridge this gap: graph neural networks pretrained on DFT and TDDFT properties can be finetuned with limited qsGW and qsGW-Bethe–Salpeter Equation (BSE) data to yield accurate predictions of quasiparticle and excitation energies. Assessing both full-model and readout-only finetuning across chemically diverse test sets, we find that pretraining improves accuracy, reduces reliance on costly qsGW data, and mitigates large predictive outliers even for molecules larger or chemically distinct from those seen during finetuning. Our results demonstrate that multi-fidelity transfer learning can substantially extend the reach of many-body-level predictions across chemical space.

1 Introduction

Accurate prediction of properties of materials and molecules underpins advances across chemical physics,[1–4] materials science,[5–7] and molecular design[8,9] for applications like optoelectronic materials[10–12] and catalysis.[13,14] Especially excited-state properties such as excitation energies are fundamental for studying processes like photosynthesis[15,16] and photovoltaic energy conversion.[10,17,18] For such purposes, computational methods are frequently employed because experiments on electronically excited states, such as determining excitation energies or characterizing short-lived molecular intermediates, are often prohibitively complex.[19–21] High-level wavefunction methods, most notably equation-of-motion coupled cluster (EOM-CC),[22,23] provide excellent descriptions of charged and neutral excitations and systematically converge to the full configuration interaction limit for weakly correlated excited states.[24–26] Yet even truncated variants such as EOM-CCSD (single and double excitations) and EOM-CCSDT (single, double and triple excitations) exhibit steep computational scaling, restricting their routine use to small molecules. Conversely, time-dependent (TD)[27,28] density functional theory (DFT)[29,30] is dramatically cheaper but also significantly less accurate.[31] Many-body perturbation theory (MBPT)[32,33] offers a more favourable accuracy-cost balance. In particular, the GW approximation[34–38] to the electronic self-energy and its combination with the Bethe–Salpeter equation (GW-BSE)[39–42] yields quasiparticle (QP) energies, and optical excitation energies with accuracy rivaling high-level wavefunction benchmarks.[43–51] Particularly the quasi-particle self-consistent variant (qsGW)[52–54] of the GW approximation frequently yields excellent results[55–58] and at the same time eliminates the dependence on the underlying mean-field reference.[59] Although being almost as efficient as TDDFT,[58] qsGW and qsGW-BSE remain computationally demanding, limiting throughput and the breadth of chemical space accessible.

Machine learning (ML) models promise to bridge this gap by providing accuracy of high-fidelity electronic-structure data at a fraction of the computational cost. Early work used kernel methods to predict GW QP energies, Green's functions, and molecular orbital levels.[60–63] Gradient boosting has been trained to correct from DFT to GW QP energies based on fingerprints encoding similarity (or dissimilarity) in energy space and in density of states projections.[64] Moving toward neural-network–based techniques, Variational Autoencoders (VAEs)[65,66] have been employed to obtain compressed latent space representations of DFT wavefunctions which are then used to predict QP energies with a Multilayer Perceptron (MLP)[67–69] and equivariant neural networks have been employed to predict Green's functions of molecules and materials.[70] In parallel, graph neural networks (GNNs) have demonstrated accurate and transferable predictions of total energies, orbital energies, and excited-state properties, enabled by sufficiently large datasets of molecular properties.[71–73] Models such as MACE,[74–76] SchNet,[77,78] DimeNet++[79] and OptiMate[80,81] have been successfully trained to predict molecular and material properties such as eigenvalue-only self-consistent GW (evGW)[82] QP energies and gaps of organic molecules[83] or the dielectric function of semiconductors and insulators.[80,81]

A major bottleneck, however, is the scarcity of high-fidelity data, *e.g.* wavefunction methods, MBPT, or experimental

Department of Chemistry and Pharmaceutical Sciences, Vrije Universiteit Amsterdam, De Boelelaan 1108, 1081 HZ Amsterdam, The Netherlands. E-mail: l.visscher@vu.nl

labels, relative to the abundance of lower-fidelity data, *e.g.* DFT and TDDFT. Large resources such as OMol25,[84] QCML[85] and QCDGE[86] provide DFT-labels for tens of millions of molecules or TDDFT-level labels for hundreds of thousands of molecules. Opposed to that, existing datasets at the wavefunction or MBPT level like OE62,[87] QM9GWBSE[88] and GDB-9-Ex_EOMCCSD[89] comprise orders of magnitude fewer samples. This disparity raises a key question: how can GNN models exploit abundant low-fidelity data while achieving high-fidelity accuracy?

Recent studies have begun to address this multi-fidelity challenge through different pretraining and finetuning strategies[90–93] and targeting, for example, pretraining on DFT or semiempirical data and finetuning, for instance, on Coupled Cluster (CC)[94–98] or Random-Phase-Approximation (RPA) labels.[99–101] Collectively, these results indicate that representations learned on low-fidelity data transfer remarkably well to high-fidelity tasks, potentially reducing the need for expensive reference calculations at the high-fidelity level.

Contributing to these efforts, we investigate whether such multi-fidelity learning can accelerate the prediction of qsGW and qsGW-BSE properties. We pretrain on DFT molecular orbital (MO) energies and TDDFT excitation energies and subsequently finetune on qsGW and qsGW-BSE labels, respectively. In contrast to prior neural network models for GW and GW-BSE which are typically trained from scratch,[10,83,102] employ Δ-learning relative to DFT,[83,103] or use descriptors from electronic structure calculations,[10,102] we leverage the multi-fidelity paradigm and provide end-to-end predictions directly from molecular structures to high-level excited-state observables. Thus, by using lower-fidelity data, we demonstrate how to achieve GW-BSE accuracy at ML computational cost, since no electronic structure calculation is needed for predictions in that way. We further demonstrate that a model pretrained on DFT and TDDFT data reduces the amount of expensive qsGW and qsGW-BSE data needed for finetuning with no loss of accuracy. Together, these results establish multi-fidelity learning as a promising path toward data-efficient surrogate models that provide MBPT accuracy at ML speed for a diverse chemical space and thus enable rapid screening of excited states properties for large sets of diverse molecules, which would not be feasible with traditional MBPT methods.

2 Methods

2.1 Data preparation

We pretrain models on two types of quantum-chemical data. For MO energies, we use up to ten million neutral molecules from the OMol25 dataset with highest occupied molecular orbital (HOMO) energies, lowest unoccupied molecular orbital (LUMO) energies and HOMO–LUMO gaps at the ωB97M-V/def2-TZVPD[104–106] level of theory, which should give reasonably close estimates of the first ionization potential (IP), the electron affinity (EA), and the fundamental gap (IP–EA) respectively. To assess the effect of data scale, we also considered one- and five-million-molecule subsets. For excitation-energy pretraining, we used TDDFT excitation energies from the QCDGE dataset at the ωB97X-D/6-31G(d)[107–109] level of theory. Finetuning was performed on qsGW QP energies and qsGW-BSE excitation energies, both taken from the QM9GWBSE dataset.

As is standard for training of neural networks, we split QM9GWBSE into training, validation, and test data. The test data serves as the first test set to which we refer to as "QM9" for brevity. Next to that, we test predictions on a subset of the PC9 (ref. 110) dataset, which matches the QM9 element space (H, C, N, O, F) and molecular size (up to 29 atoms). For the remaining two test sets, we sampled molecules from OE62 that obey either (i) the QM9 element restrictions but allow up to 48 atoms, probing extrapolation to larger systems and referred to as "OE62L", or (ii) QM9-sized molecules containing up to three heteroatoms not present in the finetuning data, probing generalization across chemical space and referred to as "OE62H". This means that we test our models on four different datasets in total. All test-set labels were recomputed with the same qsGW and qsGW-BSE settings used for the QM9GWBSE dataset. SMILES-based deduplication was applied across all datasets to prevent any overlap between pretraining, finetuning, and test molecules. Complete dataset specifications and filtering procedures are provided in the SI.

2.2 Machine learning

We employ the ViSNet architecture,[111] an SE(3)-equivariant and thus symmetry-preserving graph neural network designed to learn scalar-valued molecular properties, such as orbital energies, and, in principle, also vector-valued properties. In the present work, we restrict our use of ViSNet to scalar-valued predictions. A schematic overview of the ViSNet model, illustrated for the case of DFT pretraining followed by qsGW finetuning as performed here, is shown in Fig. 1. The model takes nuclear charges $\{Z_i\}$ and Cartesian atomic coordinates $\{r_i\}$ as input, which are embedded into initial atomic feature vectors and radial basis representations respectively. The feature vectors are subsequently updated within a series of interaction blocks *via* message passing, whereby each atomic node interacts with its local environment defined by a cutoff radius. To model these interactions, ViSNet incorporates interatomic distances, angles, dihedrals, and improper dihedrals, enabling accurate geometric representations while maintaining computational efficiency, which is essential for large-scale pretraining. In the output block, the target properties, for instance qsGW quasiparticle energies in this work, are extracted from the learned atomic feature representations. Further architectural details are provided in the original ViSNet publication.To assess the effect of model size, two model variants with different total numbers of trainable parameters are examined: a small variant with 8.9×10^5 parameters and a large variant with 2.5×10^6 parameters. Hyperparameters were selected to balance training efficiency, stability, and the accuracy of baseline (non-pretrained) models. To isolate the effect of pretraining, we maintained consistent hyperparameters between pretrained and non-pretrained runs, modifying them only when required to ensure stable convergence of training and validation losses. We consider two finetuning strategies. In the "Full" approach, all model weights are updated during finetuning. In the

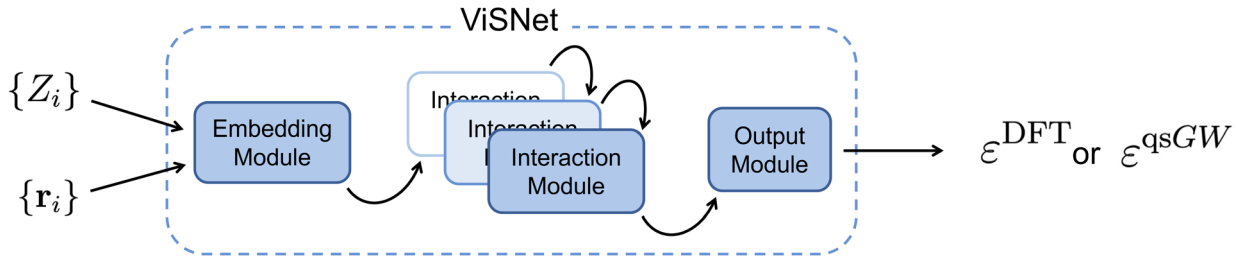

Fig. 1 Basic flowchart of the ViSNet model for predicting a single scalar-valued property, for instance a DFT orbital energy $\varepsilon^{\mathrm{DFT}}$ or a qsGW quasiparticle energy $\varepsilon^{\mathrm{qsGW}}$ based on the set of nuclear charges $\{Z_i\}$ and atomic coordinates $\{r_i\}$ of an input molecule.

"Transfer" approach, only weights in the readout layers following message-passing layers are updated, enabling reuse of latent representations learned from lower-fidelity data. Such transfer-learning strategies have proven effective for multi-fidelity molecular property prediction[90,112] and motivate their use here. Detailed hyperparameters and details on the training procedures are provided in the SI.

3 Results

We investigate whether a multi-fidelity learning paradigm can improve the accuracy of GNN-based predictions of qsGW quasiparticle (QP) energies and qsGW-BSE excitation energies. Our analysis begins with qsGW QP energies. We start by pretraining models on DFT MO energies and subsequently finetune them on qsGW QP energies. Note that the DFT pretraining data is on the range-separated hybrid DFT level which is a reasonable approximation to corresponding QP energies.[113] We then evaluate (1) whether pretraining reduces prediction errors on the test sets relative to training from scratch, thereby indicating improved generalization (2) whether pretraining enables the use of smaller finetuning datasets without sacrificing accuracy (3) whether model accuracies improve with pretraining on different but related properties, *e.g.* pretraining on HOMO energies and finetuning on HOMO–LUMO gaps. Finally, we pretrain on TDDFT excitation energies, then finetune on qsGW-BSE excitation energies and test, (4), whether analogous trends as in (1)–(3) arise for this combination of pretraining and finetuning target. All results are compared to models initialized with random weights and trained solely and directly on the target property.

3.1 Does pretraining improve generalization?

We first assess how model accuracy depends on the presence and extent of pretraining, the total number of trainable parameters (model size), and finetuning strategy. Fig. 2 compares mean absolute errors (MAEs) across all test sets for the small and large models as a function of pretraining-set size, with 0 corresponding to baselines trained from scratch. We also report purely pretrained models ("None"), without any finetuning, to isolate the effect of pretraining alone.

Models undergoing both pretraining and finetuning consistently achieve the lowest MAEs across all test sets. The effect is particularly pronounced for PC9, OE62L and OE62H,

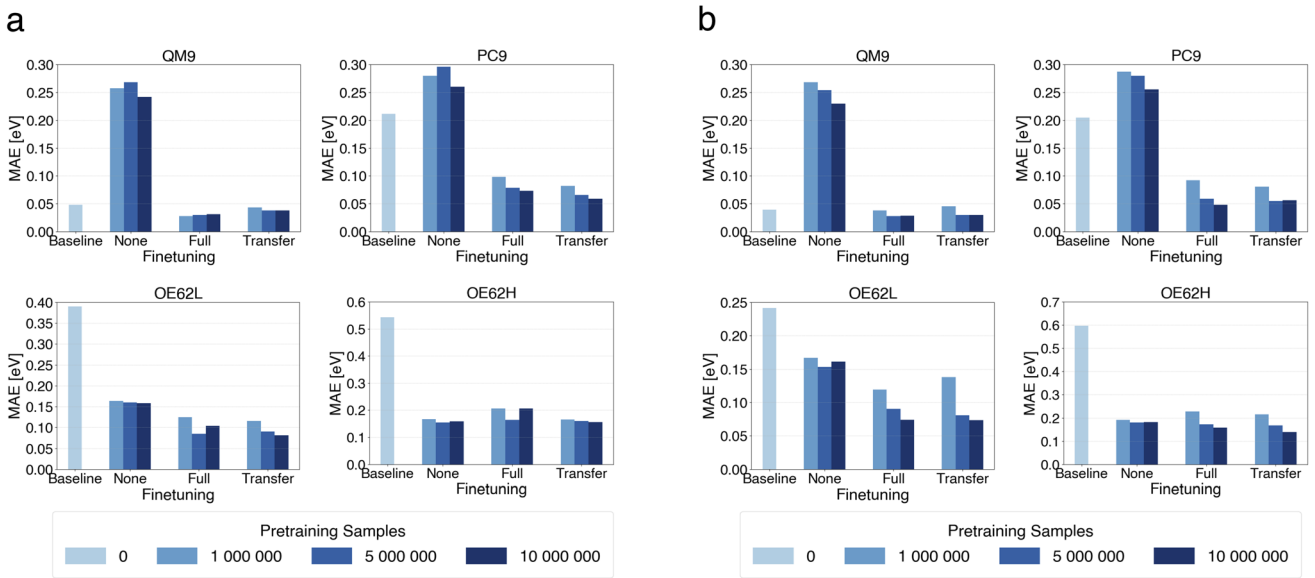

Fig. 2 MAE of qsGW QP HOMO predictions from small (a) and large (b) models pretrained on different numbers of DFT samples.

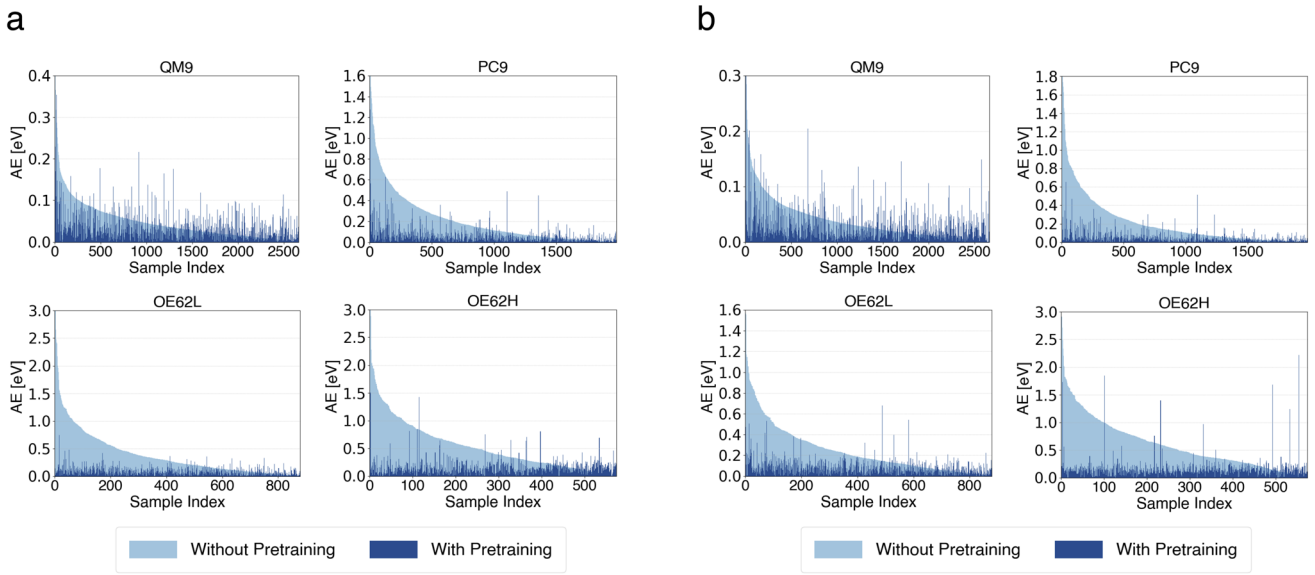

Fig. 3 Per-sample absolute errors (AEs) of qsGW QP HOMO energy predictions from small (a) and large (b) models with and without pretraining.

where reductions of up to two-thirds are observed, for instance, in the large models on OE62H. Notably, on OE62L and OE62H the purely pretrained models perform nearly as well as their finetuned counterparts, indicating that for chemically challenging cases the models rely strongly on knowledge obtained during pretraining. Nonetheless, the lowest MAEs on every set are achieved when pretraining is followed by finetuning.

Increasing model size generally yields modest accuracy gains, especially for the OE62-based test sets. Because the larger model increases the dimensionality of the feature channels, it reduces information bottlenecks in message-passing and mitigates oversquashing[114] which is especially relevant for long-range graph interactions and molecular structures with more diverse elements. However, for OE62H, both baseline models perform poorly, with almost no improvement from increased model size. In contrast, pretraining enables both small and large models to make accurate predictions despite the presence of elements unseen during finetuning on the target property.

Transfer learning performs at least as well as, and frequently better than, full finetuning for any pretraining-set size. This indicates that, after pretraining on DFT level data, only a small subset of parameters requires updating to achieve optimal accuracy on qsGW level predictions. For both model sizes, approximately 10% of the number of parameters are finetuned. This observation aligns with reports of effective transfer learning in molecular-property prediction, for instance, in drug discovery tasks.[90]

To further elucidate the source of these improvements, we examine per-sample error changes. Fig. 3 compares absolute errors for models with and without pretraining, using the small transfer-learning model with 5 000 000 pretraining samples and the large model with 10 000 000 pretraining samples. In both plots, samples are sorted in descending order by their error without pretraining. Errors after pretraining are shown in the same order such that the x-axis represents the sample index.

Across all test sets, the largest improvements occur for samples with the highest baseline errors, while slight increases in error appear for samples that were already predicted accurately. As shown in Fig. 3, the substantial reduction in errors for the most challenging samples far outweighs the modest

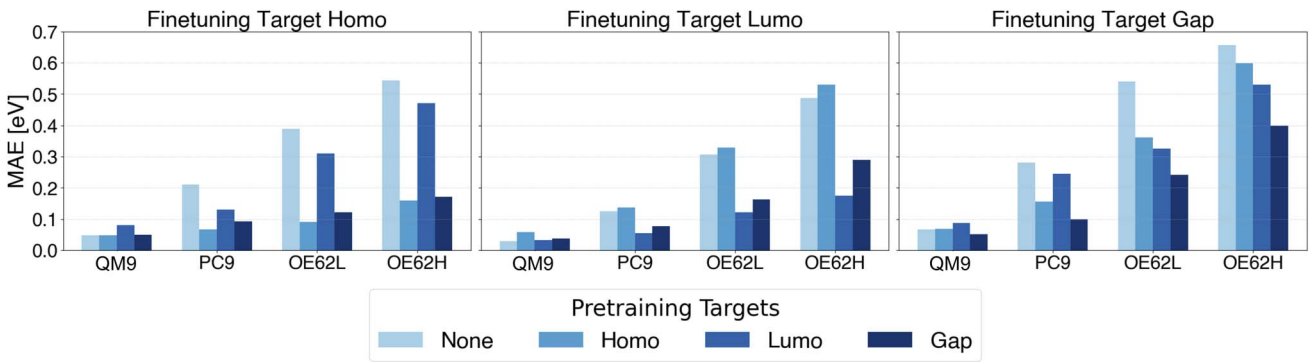

Fig. 4 MAE of QP HOMO energy, QP LUMO energy and QP gap predictions with pretraining on different QP energy targets.

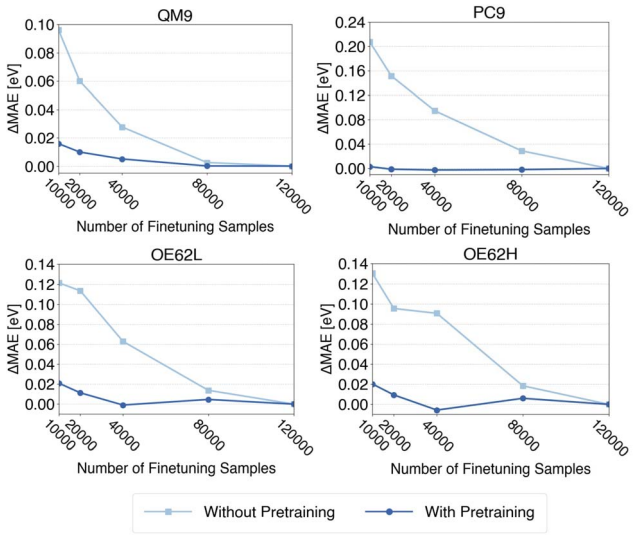

Fig. 5 MAEs of QP HOMO energy predictions after finetuning on different numbers of samples with and without prior pretraining normalized to the respective MAE when finetuning on the full finetuning-set (120 000 samples).

increases among the easiest cases, explaining the strong MAE reductions on PC9, OE62L, and OE62H. The comparatively smaller improvement on QM9 reflects the absence of large outliers for that test set. The same qualitative trends are observed for the large model. Analogous analyses for QP LUMO energies and HOMO–LUMO gaps are provided in the SI and exhibit similar qualitative behavior.

Considering the trade-off between accuracy and computational cost, small transfer-learning models with 5 000 000 pretraining samples represent an effective compromise. This combination of models and pretraining set are therefore used as the default configuration in the following unless stated otherwise.

3.2 Does pretraining reduce data demand in finetuning?

Earlier work has shown that transfer learning between low- and high-fidelity data can reduce the amount of data required for finetuning.[90,97] Here, we assess whether this holds for qsGW QP energies when pretraining on DFT MO energies. To this end, we construct finetuning sets of 10 000, 20 000, 40 000, 80 000 and, as before, the full set of 120 000 samples. Each finetuning subset is obtained by independent sampling from the full QM9GWBSE dataset rather than by incremental augmentation, ensuring maximal randomization.

Models trained with and without pretraining are evaluated on all four test sets, and the resulting learning curves for QP HOMO energies are shown in Fig. 5. Corresponding curves for QP LUMO energies and gaps are provided in the SI. Each curve is shifted by its final error (*i.e.*, the error at 120 000 samples) so that the *y*-axis reflects the deviation from the presumed minimum MAE. For every finetuning-set size, three independently sampled subsets are generated, and the reported metrics are averages over these three runs, reducing the influence of small-sample effects.

Despite averaging, models trained from scratch still show a more irregular convergence, especially for OE62L, whereas the corresponding curves for pretrained models converge noticeably smoother. This is expected: without pretraining, each finetuning sample carries greater weight, and the removal of a single informative sample can produce abrupt changes in MAE. Pretraining effectively increases the total amount of information available, thereby mitigating such fluctuations.

Most importantly, Fig. 5 shows that, across all test sets, pretrained models reach convergence with 40 000 finetuning samples, and often with only 20 000 samples, even for the most demanding cases. Thus, at most one-third of the qsGW data required for training from scratch is sufficient to achieve comparable accuracy which effectively reducing the computation cost of prodcuing high-level training data. The same trends are observed for QP LUMO energies and QP gaps (see SI).

3.3 Does pretraining transfer across properties?

Beyond reduction of test errors and qsGW data requirements, it is desirable for foundation models pretrained on one property to be reusable for finetuning on related targets. Such cross-property transfer would obviate repeated costly pretraining. To assess its feasibility, we examine all combinations of pretraining and finetuning on QP HOMO energies, QP LUMO energies, and QP gaps, and compare their performance with baselines trained directly on the target property. The resulting MAEs are shown in Fig. 4.

In nearly all cases, pretraining on any of the three properties lowers the MAE relative to the baseline. As expected, the largest improvements occur when pretraining and finetuning targets coincide. However, substantial gains also arise when the two targets are information-theoretically linked: QP gap models benefit from HOMO or LUMO pretraining, and conversely, HOMO and LUMO models benefit from gap pretraining. Because these properties are inherently related, the gap being the LUMO–HOMO difference, pretraining exposes the model to patterns directly relevant to the downstream task, yielding a more favorable initialization than random weights. In contrast, when pretraining and finetuning targets share little underlying information, the benefits vanish and can even reverse, known as negative transfer,[115,116] as reported for supervised pretraining in molecular representation learning.[117] For example, pretraining on QP HOMO energies and finetuning on QP LUMO energies increases MAEs across all test sets.

Conclusively, cross-property transfer can yield sizable error reductions even approaching the gains of perfectly aligned pretraining. Nonetheless, realizing the full benefit of pretraining still requires that both targets coincide.

3.4 Does pretraining and finetuning carry over to other excited-state targets?

Finally, we assess whether our multi-fidelity strategy extends to qsGW-BSE excitation energies. First, we test whether finetuning on qsGW-BSE benefits from pretraining on TDDFT excitation

energies. Because both methods describe neutral excitations, we expect reasonable alignment between the two targets. In all experiments, we use the lowest excitation energy per molecule. For comparison, we also evaluate models pretrained on DFT MO energies, hypothesizing that TDDFT should provide a more suitable pretraining signal than DFT. For DFT pretraining, we construct a 500 000-sample set of DFT gaps from our OMol25 subset restricted to molecules containing H, C, N, O, and F and with at most 32 atoms, mirroring the element and size limits of the QCDGE dataset used for TDDFT pretraining. This isolates the effect of the pretraining target from the effect of dataset size and diversity. We additionally pretrain a second model on the full, unconstrained 5 000 000-sample DFT dataset used earlier to probe whether increased size and chemical diversity can compensate for a less well-aligned pretraining target. The resulting MAEs are shown in Fig. 6 where the constrained DFT set is denoted "DFT constr.". Note that the *y*-axis is split because the OE62H errors are reported on a different scale.

On QM9, baseline errors are already low and finetuning has only minor impact as observed before for QP energies. On PC9 and OE62L, TDDFT pretraining clearly improves MAEs, albeit less strongly than the transfer from DFT to qsGW observed earlier. DFT pretraining on the constrained DFT set yields only marginal gains on OE62L and slightly worsens performance on PC9. The reason for that could be the misalignment between DFT gaps and qsGW-BSE excitations and is therefore in line with our previous findings about negative transfer due to insufficient alignment.

For OE62H, both TDDFT pretraining and DFT pretraining with a constrained dataset significantly increase the MAE. This is plausible, as both pretraining datasets span only a narrow element distribution, whereas OE62H contains a much more diverse range of elements. In contrast, DFT pretraining on the large, unconstrained dataset, despite its weaker target alignment, markedly reduces MAEs on OE62H presumably due to its broader coverage of chemical space. Similarly, although constrained DFT pretraining slightly increases the MAE on OE62L, the unconstrained model offers modest improvement. Notably, TDDFT pretraining still outperforms both DFT-based approaches on OE62L, underscoring its stronger alignment with qsGW-BSE. These trends suggest that larger and more chemically diverse TDDFT datasets could further lower MAEs across PC9, OE62L, and especially OE62H, potentially matching the gains observed with unconstrained DFT pretraining. Such datasets, however, are not currently available.

Second, we examine whether pretraining reduces the amount of costly qsGW-BSE data required for finetuning. Following the procedure used for QP energies, we construct finetuning sets of 10 000, 20 000, 40 000, 80 000, and the full 120 000 samples. Again, each of those subset is drawn independently from the QM9GWBSE dataset. As before, models with and without pretraining are evaluated on all four test sets. For each test set, we apply the model variant that, in the previous analysis (Fig. 6), achieved the lowest MAE when pretrained on the respective target property (DFT or TDDFT gap). Specifically, models pretrained on DFT gaps are used for PC9 and OE62H, while those pretrained on TDDFT gaps are used for QM9 and OE62L. The resulting learning curves are shown in Fig. 7. Each curve is shifted by its final error (*i.e.*, the MAE obtained with 120 000 finetuning samples), such that the *y*-axis reflects the deviation from that final error on the full finetuning set.

As in the QP energy analysis, pretraining clearly accelerates convergence of the test set errors, most notably for PC9, OE62L, and OE62H. Quantitatively, the curves indicate that approximately 40 000 samples, about one-third of the full qsGW-BSE finetuning set, are sufficient to reach convergence across all test sets. This mirrors the behavior observed for qsGW QP energies and again demonstrates that pretraining can markedly reduce the amount of high-level data required to achieve target accuracy.

In summary, the multi-fidelity paradigm extends naturally to qsGW-BSE excitation energies, reducing test errors and

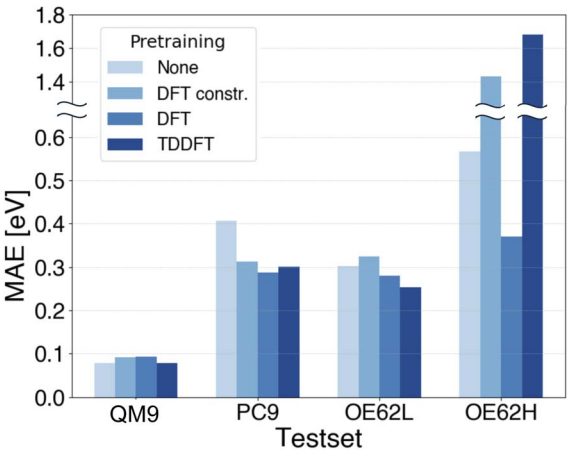

Fig. 6 MAE of qsGW-BSE excitation energy predictions after pretraining on DFT and TDDFT data.

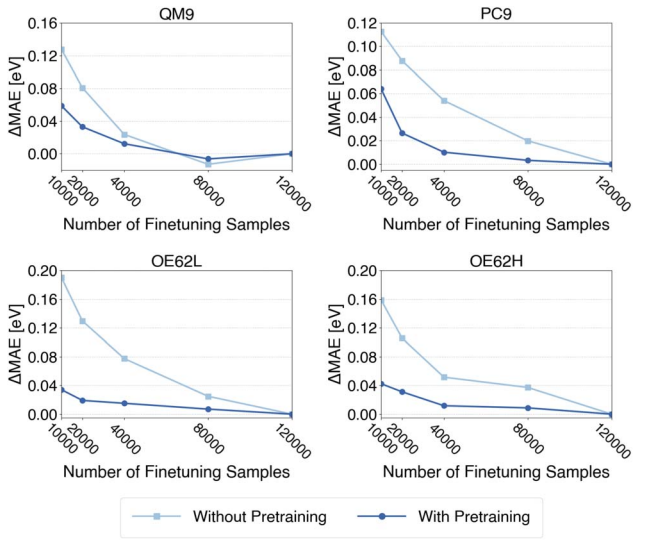

Fig. 7 MAEs of qsGW-BSE excitation energy predictions after finetuning on different numbers of samples with and without prior pretraining normalized to the respective MAE when finetuning on the full finetuning-set (120 000 samples).

reducing the computational cost of producing high-level labels for training. However, our findings also suggest that realizing the full benefits requires large pretraining datasets that are both well aligned with the target property and sufficiently broad in chemical diversity.

4 Conclusions

In this work, we investigated transfer learning from lower-fidelity DFT and TDDFT data to higher-fidelity qsGW and qsGW-BSE targets. Models were pretrained on DFT molecular orbital energies or TDDFT excitation energies and subsequently finetuned on qsGW quasiparticle energies or qsGW-BSE excitation energies, either by updating all weights or by restricting training to the readout layers following message-passing layers. The impact of these strategies was assessed across four test sets, including one containing molecules larger than those used during finetuning on the target property and another containing heteratoms also unseen during finetuning.

Our results show that pretraining on DFT- and TDDFT-level data provides a more favorable initialization of model weights than standard random initialization for learning qsGW and qsGW-BSE properties. Notably, accurate models can be obtained by leveraging only a modest subset of the abundant DFT data and by training only a fraction of the model parameters, thereby reducing cost for both data-generation and training. We observe lower test errors, driven in particular by a reduction of large outliers observed without pretraining. On top of that, we demonstrate decreased requirement for expensive qsGW and qsGW-BSE data. Specifically around 20 000 instead of 120 000 finetuning samples for qsGW and around 40 000 instead of 120 000 finetuning samples for qsGW-BSE suffice to converge the errors on our most challenging test sets. We also show evidence of the transferability of foundation models across distinct QP properties. This presents a promising opportunity to "recycle" existing foundation models rather than generating new datasets and retraining models from scratch for each target property. Thus, with our proposed strategy we lower the expected amount of data needed for both pretraining and finetuning. We emphasize that predictions made far outside the chemical space covered by our training and test sets (see the detailed composition of all datasets in the SI), for example, for molecules containing more than 250 atoms or transition metals, should be treated with caution. Although our results demonstrate that DFT pretraining can substantially reduce prediction errors for qsGW quasiparticle energies and GW-BSE excitation energies within a given chemical subspace and even beyond it, this study does not establish that such extrapolation reaches undefinitely far into all chemical space. Likewise, predictions for systems exhibiting pronounced multi-reference character should be handled with caution, as our models are pretrained on DFT and TDDFT data and finetuned exclusively on single-reference systems.

Overall, this study demonstrates that data-efficient GNN models capable of end-to-end prediction at the level of MBPT can be realized, with generalization extending into regions of chemical space not encountered during finetuning. At the same time, our findings underscore key challenges, including the need for sufficiently large and chemically diverse low-fidelity datasets. We show that the careful selection of pretraining targets of adequate level of theory that sufficiently align with finetuning targets are crucial to fully unlock the benefits of pretraining. Future work could explore whether the strategy of DFT or TDDFT pretraining followed by GW-BSE finetuning can be extended to additional target quantities, including other scalar properties such as oscillator strengths, as well as vector-valued properties such as transition dipoles.

Author contributions

D. B. conceptualized the project, conducted the calculations and analyzed the data. A. F. and L. V. provided guidance and supervision throughout the project. L. V. acquired funding that supported this work. D. B wrote the manuscript which was reviewed by A. F. and L. V.

Conflicts of interest

There are no conflicts to declare.

Data availability

Code, test data and model checkpoints for this work are publicly available at https://github.com/daoiradrio/ViSNetGWBSE. The DFT pretraining data is taken from https://huggingface.co/facebook/OMol25/blob/main/DATASET.md. The TDDFT pretraining data is taken from https://langroup.site/QCDGE/. The qsGW and qsGW-BSE finetuning data is taken from https://zenodo.org/records/17902233.

Supplementary information (SI): details on test set compositions, model hyperparameters, training parameters and additional analyses like error bars and learning curves. See DOI: https://doi.org/10.1039/d5sc09780k.

Acknowledgements

We acknowledge the use of supercomputer facilities at SURFsara sponsored by NWO Physical Sciences, with financial support from The Netherlands Organization for Scientific Research (NWO). LV and DB acknowledge funding from Microsoft Research. AF acknowledges funding through a VENI grant from NWO under grant agreement VI. Veni.232.013. The authors thank Ansgar Pausch for contributions and suggestions in the initial ideation phase.

References

1 T. Froitzheim, M. Müller, A. Hansen and S. Grimme, *J. Chem. Phys.*, 2025, **162**, 214109.
2 M. Müller, T. Froitzheim, A. Hansen and S. Grimme, *J. Phys. Chem.*, 2024, **128**, 10723–10736.
3 L. A. Rosset and V. L. Deringer, *arXiv*, 2025, preprint arXiv:2510.15633, DOI: 10.48550/arXiv.2510.15633.

4 Y. Zhou, D. F. Thomas du Toit, S. R. Elliott, W. Zhang and V. L. Deringer, *Nat. Commun.*, 2025, **16**, 8688.
5 K. Ueltzen, A. A. Naik, C. Ertural, P. Benner and J. George, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-xj84d.
6 Y. Zimmermann, A. Bazgir, A. Al-Feghali, M. Ansari, J. Bocarsly, L. C. Brinson, Y. Chiang, D. Circi, M.-H. Chiu, N. Daelman, *et al.*, *Mach. Learn.: Sci. Technol.*, 2025, **6**, 030701.
7 A. A. Naik, C. Ertural, N. Dhamrait, P. Benner and J. George, *Sci. Data*, 2023, **10**, 610.
8 B. Szabó, P. Zaby, L. Dick, K. Drysch, Y. Dawer, W. Reckien, A. Udvardy, F. Neese, B. Kirchner and O. Hollóczki, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-m5hnw.
9 L. Dick and B. Kirchner, *J. Chem. Inf. Model.*, 2023, **63**, 6706–6716.
10 T. Biswas, A. Gupta and A. K. Singh, *RSC Adv.*, 2025, **15**, 8253–8261.
11 M. Grunert, M. Großmann and E. Runge, *Phys. Rev. B*, 2024, **110**, 075204.
12 W. Barford, J. L. Gardner and J. R. Mannouch, *Faraday Discuss.*, 2020, **221**, 281–298.
13 E. Keller, V. Blum, K. Reuter and J. T. Margraf, *J. Chem. Phys.*, 2025, **162**, 074111.
14 S. Stocker, H. Jung, G. Csányi, C. F. Goldsmith, K. Reuter and J. T. Margraf, *J. Chem. Theory Comput.*, 2023, **19**, 6796–6804.
15 Y.-C. Cheng and G. R. Fleming, *Annu. Rev. Phys. Chem.*, 2009, **60**, 241–262.
16 C. Curutchet and B. Mennucci, *Chem. Rev.*, 2017, **117**, 294–343.
17 M. Gruber, J. Wagner, K. Klein, U. Hörmann, A. Opitz, M. Stutzmann and W. Brütting, *Adv. Energy Mater.*, 2012, **2**, 1100–1108.
18 O. V. Mikhnenko, P. W. Blom and T.-Q. Nguyen, *Energy Environ. Sci.*, 2015, **8**, 1867–1888.
19 J. S. Lim and S. K. Kim, *Nat. Chem.*, 2010, **2**, 627–632.
20 M. Jia, J. Kong, H. Zhou, J. Chen, S. Zhang and M. Zhou, *Laser-based Techniques for Nanomaterials: Processing to Characterization*, 2024, pp. 262–286, DOI: 10.1039/9781837673513-00262.
21 R. Geneaux, H. J. Marroux, A. Guggenmos, D. M. Neumark and S. R. Leone, *Philos. Trans. R. Soc. A*, 2019, **377**, 20170463.
22 H. J. Monkhorst, *Int. J. Quantum Chem.*, 1977, **12**, 421–432.
23 J. F. Stanton and R. J. Bartlett, *J. Chem. Phys.*, 1993, **98**, 7029–7039.
24 P.-F. Loos, A. Scemama, A. Blondel, Y. Garniron, M. Caffarel and D. Jacquemin, *J. Chem. Theor. Comput.*, 2018, **14**, 4360–4379.
25 P.-F. Loos, A. Scemama and D. Jacquemin, *J. Phys. Chem. Lett.*, 2020, **11**, 2374–2383.
26 A. Marie and P.-F. Loos, *J. Chem. Theory Comput.*, 2024, **20**, 4751–4777.
27 E. Runge and E. K. Gross, *Phys. Rev. Lett.*, 1984, **52**, 997.
28 M. Petersilka, U. Gossmann and E. Gross, *Phys. Rev. Lett.*, 1996, **76**, 1212.
29 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864.
30 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133.
31 P.-F. Loos and D. Jacquemin, *J. Phys. Chem.*, 2021, **125**, 10174–10188.
32 A. Fetter and J. Walecka, Mineola, NY, 2003.
33 R. M. Martin, L. Reining and D. M. Ceperley, *Interacting electrons*, Cambridge University Press, 2016.
34 L. Hedin, *Phys. Rev.*, 1965, **139**, A796.
35 F. Aryasetiawan and O. Gunnarsson, *Rep. Prog. Phys.*, 1998, **61**, 237.
36 L. Reining, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2018, **8**, e1344.
37 D. Golze, M. Dvorak and P. Rinke, *Front. Chem.*, 2019, **7**, 377.
38 A. Marie, A. Ammar and P. F. Loos, *Adv. Quantum Chem.*, 2024, **90**, 157–184.
39 E. E. Salpeter and H. A. Bethe, *Phys. Rev.*, 1951, **84**, 1232.
40 M. Gell-Mann and F. Low, *Phys. Rev.*, 1951, **84**, 350.
41 G. Strinati, *Riv. Nuovo Cimento*, 1988, **11**, 1–86.
42 G. Onida, L. Reining and A. Rubio, *Rev. Mod. Phys.*, 2002, **74**, 601.
43 F. Bruneval and A. Forster, *J. Chem. Theory Comput.*, 2024, **20**, 3218–3230.
44 J. W. Knight, X. Wang, L. Gallandi, O. Dolgounitcheva, X. Ren, J. V. Ortiz, P. Rinke, T. Körzdörfer and N. Marom, *J. Chem. Theor. Comput.*, 2016, **12**, 615–626.
45 F. Caruso, M. Dauth, M. J. Van Setten and P. Rinke, *J. Chem. Theor. Comput.*, 2016, **12**, 5076–5087.
46 F. Bruneval, N. Dattani and M. J. van Setten, *Front. Chem.*, 2021, **9**, 749779.
47 D. Golze, L. Keller and P. Rinke, *J. Phys. Chem. Lett.*, 2020, **11**, 1840–1847.
48 J. Li, Y. Jin, P. Rinke, W. Yang and D. Golze, *J. Chem. Theory Comput.*, 2022, **18**, 7570–7585.
49 C. A. McKeon, S. M. Hamed, F. Bruneval and J. B. Neaton, *J. Chem. Phys.*, 2022, **157**, 074103.
50 D. Jacquemin, I. Duchemin and X. Blase, *J. Phys. Chem. Lett.*, 2017, **8**, 1524–1529.
51 I. Knysh, F. Lipparini, A. Blondel, I. Duchemin, X. Blase, P.-F. Loos and D. Jacquemin, *J. Chem. Theory Comput.*, 2024, **20**, 8152–8174.
52 S. V. Faleev, M. van Schilfgaarde and T. Kotani, *Phys. Rev. Lett.*, 2004, **93**, 126406.
53 M. van Schilfgaarde, T. Kotani and S. Faleev, *Phys. Rev. Lett.*, 2006, **96**, 226402.
54 T. Kotani, M. Van Schilfgaarde and S. V. Faleev, *Phys. Rev. B Condens. Matter*, 2007, **76**, 165106.
55 A. Förster and L. Visscher, *Phys. Rev. B*, 2022, **105**, 125121.
56 A. Forster, *J. Chem. Theor. Comput.*, 2025, **21**, 1709–1721.
57 I. Duchemin and X. Blase, *J. Chem. Phys.*, 2025, **162**, 054121.
58 A. Förster and L. Visscher, *J. Chem. Theor. Comput.*, 2022, **18**, 6779–6793.
59 A. Förster and L. Visscher, *Front. Chem.*, 2021, **9**, 736591.
60 O. Caylak and B. Baumeier, *J. Chem. Theory Comput.*, 2021, **17**, 4891–4900.

61 C. Venturella, C. Hillenbrand, J. Li and T. Zhu, *J. Chem. Theory Comput.*, 2023, **20**, 143–154.
62 C. Venturella, J. Li, C. Hillenbrand, X. Leyva Peralta, J. Liu and T. Zhu, *Nat. Comput. Sci.*, 2025, 1–12.
63 A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen and P. Rinke, *J. Chem. Phys.*, 2019, **150**, 204121.
64 N. R. Knøsgaard and K. S. Thygesen, *Nat. Commun.*, 2022, **13**, 468.
65 D. P. Kingma and M. Welling, *arXiv*, 2013, preprint arXiv:1312.6114, DOI: 10.48550/arXiv.1312.6114.
66 D. J. Rezende, S. Mohamed and D. Wierstra, *International conference on machine learning*, 2014, pp. 1278–1286.
67 B. Hou, J. Wu and D. Y. Qiu, *Nat. Commun.*, 2024, **15**, 9481.
68 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *nature*, 1986, **323**, 533–536.
69 I. Goodfellow, *Deep learning*, 2016.
70 X. Dong, E. Gull and L. Wang, *Phys. Rev. B*, 2024, **109**, 075112.
71 A. Fediai, P. Reiser, J. E. O. Peña, P. Friederich and W. Wenzel, *Sci. Data*, 2023, **10**, 581.
72 L. Ruddigkeit, R. Van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem. Inf. Model.*, 2012, **52**, 2864–2875.
73 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, *Sci. Data*, 2014, **1**, 1–7.
74 I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, *et al.*, *J. Chem. Phys.*, 2025, **163**, 184110.
75 I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csányi, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 11423–11436.
76 I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. Simm, R. Drautz, C. Ortner, B. Kozinsky and G. Csányi, *Nat. Mach. Intell.*, 2025, **7**, 56–67.
77 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.
78 K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko and K.-R. Müller, *Adv. Neural Inf. Process. Syst.*, 2017, **30**, 992–1002.
79 J. Gasteiger, S. Giri, J. T. Margraf and S. Günnemann, *arXiv*, 2020, preprint arXiv:2011.14115, DOI: 10.48550/arXiv.2011.14115.
80 M. Grunert, M. Großmann and E. Runge, *Phys. Rev. Mater.*, 2024, **8**, L122201.
81 M. Grunert, M. Großmann and E. Runge, *Small*, 2025, 2412519.
82 M. P. Surh, S. G. Louie and M. L. Cohen, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 1991, **43**, 9126.
83 A. Fediai, P. Reiser, J. E. O. Peña, W. Wenzel and P. Friederich, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 035045.
84 D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba, P. Eastman *et al.*, *arXiv*, 2025, preprint arXiv:2505.08762, DOI: 10.48550/arXiv.2505.08762.
85 S. Ganscha, O. T. Unke, D. Ahlin, H. Maennel, S. Kashubin and K.-R. Müller, *Sci. Data*, 2025, **12**, 406.
86 Y. Zhu, M. Li, C. Xu and Z. Lan, *Sci. Data*, 2024, **11**, 948.
87 A. Stuke, C. Kunkel, D. Golze, M. Todorović, J. T. Margraf, K. Reuter, P. Rinke and H. Oberhofer, *Sci. Data*, 2020, **7**, 58.
88 D. Baum, A. Förster and L. Visscher, qsGW quasiparticle and GW-BSE excitation energies of 133,885 molecules, 2025, https://arxiv.org/abs/2512.10815.
89 K. Mehta, M. L. Pasini, D. Ganyushin, P. Yoo and S. Irle, *IEEE Data Descr.*, 2025.
90 D. Buterez, J. P. Janet, S. J. Kiddle, D. Oglic and P. Lió, *Nat. Commun.*, 2024, **15**, 1517.
91 M. Grunert, M. Großmann, J. Hänseroth, A. Flötotto, J. Oumard, J. L. Wolf, E. Runge and C. Dreßler, *J. Phys. Chem. C*, 2025, **129**, 9662–9669.
92 J. L. Gardner, K. T. Baker and V. L. Deringer, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 015003.
93 J. L. Gardner, D. F. Toit, C. B. Mahmoud, Z. F. Beaulieu, V. Juraskova, L.-B. Paşca, L. A. Rosset, F. Duarte, F. Martelli, C. J. Pickard *et al.*, *arXiv*, preprint arXiv:2506.10956, DOI: 10.48550/arXiv.2506.109562025.
94 K. Raghavachari, G. W. Trucks, J. A. Pople and M. Head-Gordon, *Chem. Phys. Lett.*, 1989, **157**, 479–483.
95 I. Purvis, G. D. and R. J. Bartlett, *J. Chem. Phys.*, 1982, **76**, 1910–1918.
96 J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev and A. E. Roitberg, *Nat. Commun.*, 2019, **10**, 2903.
97 V. Zaverkin, D. Holzmüller, L. Bonfirraro and J. Kästner, *Phys. Chem. Chem. Phys.*, 2023, **25**, 5383–5396.
98 J. L. Gardner, H. Schulz, J. Helie, L. Sun and G. N. Simm, *arXiv*, 2025, preprint arXiv:2506.14963, DOI: 10.48550/arXiv.2506.14963.
99 D. Bohm and D. Pines, *Phys. Rev.*, 1951, **82**, 625.
100 M. Grunert, M. Großmann and E. Runge, *Nat. Commun.*, 2025, **16**, 8142.
101 M. Cui, K. Reuter and J. T. Margraf, *Mach. Learn.: Sci. Technol.*, 2025, **6**, 015071.
102 B. Hou, X. Xu, J. Wu and D. Y. Qiu, *arXiv*, 2025preprint arXiv:2507.05480, DOI: 10.48550/arXiv.2507.05480.
103 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, *J. Chem. Theor. Comput.*, 2015, **11**, 2087–2096.
104 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2016, **144**, 214110.
105 O. A. Vydrov and T. Van Voorhis, *J. Chem. Phys.*, 2010, **133**, 244103.
106 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
107 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.
108 W. J. Hehre, R. Ditchfield and J. A. Pople, *J. Chem. Phys.*, 1972, **56**, 2257–2261.
109 P. C. Hariharan and J. A. Pople, *Theor. Chim. Acta*, 1973, **28**, 213–222.
110 M. Glavatskikh, J. Leguy, G. Hunault, T. Cauchy and B. Da Mota, *J. Cheminf.*, 2019, **11**, 1–15.
111 Y. Wang, T. Wang, S. Li, X. He, M. Li, Z. Wang, N. Zheng, B. Shao and T.-Y. Liu, *Nat. Commun.*, 2024, **15**, 313.
112 M. Radova, W. G. Stark, C. S. Allen, R. J. Maurer and A. P. Bartók, *npj Comput. Mater.*, 2025, **11**, 237.

113 I. C. Gerber and J. G. Angyán, *Chem. Phys. Lett.*, 2005, **415**, 100–105.
114 U. Alon and E. Yahav, *arXiv*, 2020, preprint arXiv:2006.05205, DOI: 10.48550/arXiv.2006.05205.
115 Z. Wang, Z. Dai, B. Póczos and J. Carbonell, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11293–11302.
116 W. Zhang, L. Deng, L. Zhang and D. Wu, *IEEE/CAA J. Autom. Sinica.*, 2022, **10**, 305–329.
117 R. Sun, H. Dai and A. W. Yu, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 12096–12109.