Journal of Cheminformatics

RESEARCH Open Access

Pretraining graph transformers with atom-in-a-molecule quantum properties for improved ADMET modeling

Alessio Fallani[1,2], Ramil Nugmanov[2*], Jose Arjona-Medina[2], Jörg Kurt Wegner[3], Alexandre Tkatchenko[1] and Kostiantyn Chernichenko[2*]

Abstract

We evaluate the impact of pretraining Graph Transformer architectures on atom-level quantum-mechanical features for the modeling of absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of drug-like compounds. We compare this pretraining strategy with two others: one based on molecular quantum properties (specifically the HOMO-LUMO gap) and one using a self-supervised atom masking technique. After fine-tuning on Therapeutic Data Commons ADMET datasets, we evaluate the performance improvement in the different models observing that models pretrained with atomic quantum mechanical properties produce in general better results. We then analyze the latent representations and observe that the supervised strategies preserve the pretraining information after fine-tuning and that different pretrainings produce different trends in latent expressivity across layers. Furthermore, we find that models pretrained on atomic quantum mechanical properties capture more low-frequency Laplacian eigenmodes of the input graph via the attention weights and produce better representations of atomic environments within the molecule. Application of the analysis to a much larger non-public dataset for microsomal clearance illustrates generalizability of the studied indicators. In this case the performances of the models are in accordance with the representation analysis and highlight, especially for the case of masking pretraining and atom-level quantum property pretraining, how model types with similar performance on public benchmarks can have different performances on large scale pharmaceutical data.

Scientific contribution

We systematically compared three different data type/methodologies for pretraining molecular Graphormer with the purpose of modeling ADMET properties as downstream tasks. The learned representations from differently pretrained models were analyzed in addition to comparison of downstream task performances that have been typically reported in similar works. Such examination methodologies, including a newly introduced analysis of Graphormer's Attention Rollout Matrix, can guide pretraining strategy selection, as corroborated by a performance evaluation on a larger internal dataset.

*Correspondence:
Ramil Nugmanov
rnugmano@its.jnj.com
Kostiantyn Chernichenko
kcherni1@its.jnj.com
Full list of author information is available at the end of the article

Introduction

Effectively representing molecules for modeling applications is a fundamental challenge in cheminformatics and machine learning, which lead to the development of various representation methods. In the realm of precomputed representations, different approaches are utilized depending on the available input data. Fingerprints, which encode the presence or absence of substructures and certain chemical properties in binary vectors, are commonly used in cheminformatics, particularly when 3D data is not available. On the other hand, physics-inspired representations like the Coulomb Matrix [1], Bag of Bonds[2], SLATM [3] and many others are more frequently employed for representing 3D geometries and physical properties. With the advent of deep learning, the potential to learn representations directly from data has become increasingly apparent[4–9]. These learned representations, shaped by the training of deep neural networks, enable the transformation of input data into a latent space where relevant features can be distilled in different ways for specific tasks[10–12]. For instance, contrastive learning techniques have been employed to learn representations including information from other modalities such as image information[13] and knowledge-graph information[14]. Similarly, other learning approaches have been developed to integrate multimodal data with chemical structures, such as medical records [15], natural language [16] or pooling together sequence graph and geometry information [17]. Moreover, this approach often allows for the incorporation of invariance or equivariance with respect to particular transformations, enhancing the robustness and accuracy of the models [18–20]. However, despite the remarkable successes achieved, these methodologies still present some limitations that need to be carefully evaluated [21]. In particular, challenges such as data scarcity[22] and generalizability remain pertinent concerns in the field [9, 23–27]. To address these challenges, the concept of pretraining models on related tasks or employing self-supervised learning strategies has gained significant traction. The success of this methodology is evident, for example, in the realm of natural language processing, where overparametrized large language models (LLMs) are pretrained on a wide corpus of data, and then made available for fine-tuning with minimal resources and small datasets on specific tasks [28, 29]. Following a similar paradigm, in the context of molecular representation learning this technique has been explored as a mean to enhance model generalizability and performance across various downstream tasks [11, 30–33]. The selection of pretraining data and tasks, though, is not trivial. The data should be such that: (i) it is available or can be generated at scale and (ii) provide fundamental information about molecular properties and behavior. A natural choice following these criteria is quantum mechanical (QM) reference data, as it is known to be related to fundamental aspects of molecular behavior [34–37] with profound implications in biochemical research and it only requires computational resources for production at scale, being in fact already present in an increasing number of public datasets [38–49]. Studies utilizing both atomic and molecular QM properties for pretraining have been already carried out in multiple excellent works [50–55] using different molecular representations and architectures. In these studies some degree of improvement on various downstream tasks is shown, but the conclusions are often based solely on the modeling benchmark data and explained by knowledge transfer between tasks, limiting the understanding of the real impact that different pretraining methods can have on the representations learned by the models. In this context, our study aims to bridge this gap and proposes a series of studies focused on investigating the impact that pretraining on atom-level quantum-mechanical (QM) properties has on the representation learned by a Graphormer neural network [56] when compared to other commonly employed pretraining strategies. The evaluation is carried out on public benchmark data and multiple analysis are performed on fine-tuned models with the addition of a test on internal microsomal clearance data. The results show that models pretrained on atom-level QM properties result in better representations under multiple indicators, and that the ranking based on those indicators matches better with the results on the larger internal dataset rather than with the results on the benchmark.

Methods

We consider a custom implementation of Graphormer [56, 57] as an instance of network that belongs to the increasingly popular family of Graph Transformers (GTs) [58], models that generally utilize a transformer architecture on 2D graph input data. As a comparison for the models pretrained on atomic QM properties, besides training the model without pretraining, we consider masking pretraining (atom-level self-supervised method) using the same dataset employed for the atom-resolved QM properties [59], and pretraining on on a much bigger dataset of a molecular property, HOMO-LUMO gap (HLG) calculated by QM methods [60]. This choice is dictated by the approximate matching of the overall number of atomic properties (molecules times non-hydrogen atoms) with the number of data points for the HLG (one per molecule). The pretrainings were followed by fine-tuning on individual target downstream tasks

from absorption, distribution, metabolism, excretion, and toxicity (ADMET) benchmark datasets of the Therapeutics Data Commons (TDC) [61]. They represent the key properties relevant to pharmacokinetics and pharmacodynamics of drugs. Other than comparing the results of these different pretraining strategies on the benchmark metrics, the studies carried out in this work investigate multiple aspects of learned latent representations. Namely, we evaluate the conservation of the pretraining information after fine-tuning, analyze expressivity of the latent representation across layers and sensitivity of the receptive field of the obtained atomic representations. We also propose and perform a novel spectral analysis of the Attention Rollout matrix [62], that studies its relation to the graph-Laplacian eigenmodes of the input molecule. Furthermore, the pretraining methodologies were utilized in the modeling of an internal company dataset of microsomal clearance (which contains much more data than its public TDC counterpart) that revealed the limitations of using only public benchmark metrics for methodology evaluation. In this section we will describe in detail the model, the datasets, the methods used for pretraining and fine-tuning as well as each of the analyses done on the fine-tuned models.

Model description

Graphormer is a GT where the input molecule is seen as a graph with atoms as nodes and bonds as edges. This model in general works by encoding the atoms in the molecule as tokens based on their atom type with an additional token (the CLS token) to encode global information about the molecule as in BERT [63], and then repeatedly applying transformer encoder layers with modified self-attention blocks with an internal bias term before the softmax function. This term is based on the topological distance matrix of the molecular graph and allows the encoding of the structural information of the molecular graph. In particular, the network employed in this work is an implementation of Graphormer from [57], inspired by the work [56]. In this implementation the centrality encoder is adapted from using only explicit neighbors to including both explicit atoms and implicit hydrogens. As a result of the combination of this modified centrality encoding together with the usual atom type encoder, the hybridization of atoms is handled implicitly. For this reason this implementation does not present any edge encoder component. For what concerns the choice of hyperparameters, we did not run hyperparameter tuning experiments as absolute performance is not the focus of this work. We purposely chose 20 hidden layers, a higher number than usually found in similar architectures, while maintaining a number of parameters that is comparable with other Transformer-based implementations previously introduced[56, 64] ($\sim$ 10M parameters). This choice is done in order to study the effects of the pretraining strategies on very deep models, considering quantities closely related to known depth-related phenomenons in machine learning literature [58, 65, 66], while maintaining reasonable training times and a number of parameters comparable to other models. The rest of the hyperparameters were chosen based on our experience and maintaining the same conditions across all models in both pretraining and fine-tuning stages for fairness of comparison. Finally, differently from our preliminary results in [67, 68] we do not employ task specific virtual nodes but rather rely solely on the original implementation in [57]. This last choice is made to exclude any effect that this technique may have on the final results, especially considering that other GT architectures generally do not employ it. More information on the Graphormer implementation is reported in the SI.

Pretraining datasets and methodology

For pretraining, we used a publicly available dataset [59] consisting of $\sim$ 136k organic molecules for a total of over 2M heavy atoms. Each molecule is represented by a single conformer initially generated using the Merck Molecular Force Field (MMFF94s) in RDKit library. The geometry for the lowest-lying conformer was then optimized at the GFN2-xtb level of theory followed by refinement of the electronic structure with DFT (B3LYP/def2svp). The dataset reports several atomic properties: a charge, electrophilic and nucleophilic Fukui indexes, an NMR shielding constant. Another pretraining dataset, PCQM4Mv2, consists of a single molecular property per molecule, an HLG that was also calculated using quantum chemistry methods https://ogb.stanford.edu/docs/lsc/pcqm4mv2/. The dataset contains over 2M of molecules and was curated under the PubChemQC project [60]. It is important to specify that albeit both datasets also contain the 3D molecular geometries, we only employ the 2D graph chemical structures.

The pretraining on atom-level QM properties is achieved via a regression task by applying a linear layer to the obtained node representations, each corresponding to a heavy (non-hydrogen) atom. The model is trained on each one of the available atomic properties separately, as well as on all of them at the same time in a multi-task setting. As a result, we obtain from these pretraining efforts 5 different models.

Pretraining on molecular quantum properties is achieved via a regression task on the values of HLG from the PCQM4Mv2 dataset and where the output

is obtained by applying a linear layer to the class token representation at the last layer of the network.

Masking pretraining, instead, is carried out in a similar way to what is usually done in BERT-based models [63, 69]. This procedure entails randomly masking 15% of the input graph node tokens by replacing them with the mask token, and then training the model to restore the correct node type from the masked embedding as a multi-class classification task. This last pretraining is carried out on the molecular structures present in the dataset used for atomic QM properties.

Downstream tasks

For the benchmarking of the obtained pretrained models, we used the absorption, distribution, metabolism, excretion, and toxicity (ADMET) group of the TDC dataset, consisting of 9 regression and 13 binary classification tasks for modeling biochemical molecular properties https://tdcommons.ai/benchmark/admet_group/overview/. The training and testing on this dataset is carried out in the same way as any molecular property modeling. For splittings and evaluation metrics we follow the guidelines of the benchmark group that we consider, hence we refer to [61]. All the downstream task trainings followed the same procedure with weights taken from the respective pretrained models (without freezing any layer) or randomly initialized for training the scratch model. No multitask training is adopted here and a different model is obtained for each split of each downstream task. For each combination of downstream task and pretraining, we obtained 5 models, corresponding to training/validation splits as provided in the benchmark, and reported the final performance as mean and standard deviation over this set. In summary, the non-pretrained Graphormer version used as a baseline model was compared with 7 different pretrained models: one per each of the 4 atom-resolved QM properties (atomic charges, NMR shielding constants, electrophilic and nucleophilic Fukui function indexes), one pretrained on all atomic properties in a multi-task setting, one for the molecule-level property (HLG), and one for masking node pretraining. Each of the 8 models is then fine-tuned on the 22 tasks from the TDC ADMET benchmark totaling 880 final models with the default 5 train/test splits per each task.

Finally, we consider a much larger set of proprietary JNJ data containing values of human liver microsome (HLM) intrinsic clearance of $\sim$130k compounds measured in two different assays. Fine-tuning on this dataset was conducted in the same way as in the case of the regression tasks in TDC with the difference that two values of clearance were modeled using a multi-task approach. A test set accounts for 20% of a total dataset size and is obtained as a scaffold split that maximizes Tanimoto distance between the train and test splits. The results are reported both in terms of R^2 coefficient and Spearman's correlation coefficient as mean and standard deviation over 3 seeds.

The mutual distribution of chemical spaces covered by pretraining and downstream datasets was analyzed using Uniform Manifold Approximation and Projection (UMAP) plot[70] of the respective Morgan fingerprints (Fig. 1). This analysis demonstrated significant overlap between the chemical spaces of the HLG and QM136 datasets on one hand and the HLM and TDC datasets on the other hand. Also, there is a partial overlap between the TDC, HLG and QM136 datasets.

Conservation of pretraining information after fine-tuning

In order to understand if the fine-tuned models preserve some of the information learned during the supervised pretraining stages or if that amounts only to a different network initialization, we analyze the latent representation obtained in the last layer. In particular, for each fine-tuning task and for each set of differently pretrained models, we freeze the model obtained from one of the seeds and encode a sample of 5000 molecules from each of the two pretraining datasets. The latent representations are split into equal size train/test sets and fit with the regularized linear regressor from [71] to reveal to which extent the representation still preserve linear correlation with the pretraining labels. The results are reported in terms of R^2 coefficient over the test set averaged across the 22 fine-tuned models. We perform this analysis in an all-to-all fashion: namely for every model against every pretraining task and also considering the models trained from scratch as some correlation may arise from learning representations during the downstream tasks. Using similar methodology, we also analyzed correlation between the representations and pretraining labels prior to fine-tuning to obtain reference points for understanding possible degradation of pretraining information.

Latent expressivity across layers

For deep transformer models and deep graph neural networks, a tendency for the latent representations of each of the considered tokens or nodes to become increasingly similar as more and more layers are applied is quite common. While a more detailed analysis of this phenomenon, as testified by the multiple studies around this topic [72–76], would require a specialized work and goes beyond the scope of this paper, characterizing this effect is of particular interest for machine learning practitioners dealing with deep models. To this end, the internal representation at each layer of the models

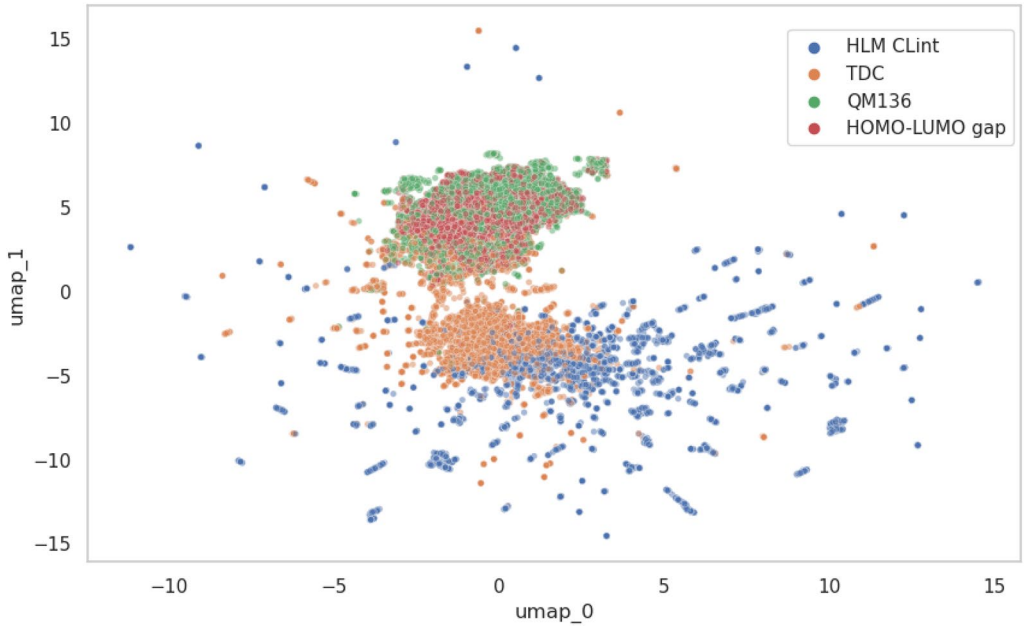

Fig. 1 UMAP plot of chemical spaces covered by the studied datasets

is also studied by analyzing a quantity introduced in [65], which is related to representation rank, that measures how similar are latent token representations. If $\mathrm{GT_L(X)} \in \mathbb{R}^{\mathrm{n\times d}}$ is the latent representation of an encoded input $\mathbf{X} \in \mathbb{R}^{n\times d}$ at layer L of a GT network, this is defined as:

$$\rho_L = \frac{||\mathrm{res}(\mathrm{GT_L(X)})||_{1,\infty}}{||\mathrm{GT_L(X)}||_{1,\infty}} \tag{1}$$

with $||(\cdot)||_{1,\infty} = \sqrt{||(\cdot)||_1||(\cdot)||_\infty}$, where $\mathrm{res}(\mathbf{X}) = \mathbf{X} - \mathbf{1}\mathbf{x}^T$, with $\mathbf{x} = \underset{\mathbf{x}}{\mathrm{argmin}}||\mathbf{X} - \mathbf{1}\mathbf{x}^T||$ where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{1} \in \mathbb{R}^n$. Namely, this metric measures how close is the representation to the closest representation in $||(\cdot)||$ norm where all n latent representations are equal to the same vector $\mathbf{x}$. We report the value of this quantity across all layers for every model, computed using a random sample of 100 molecules from each test set of the ADMET tasks.

Spectral analysis of attention rollout

To have a better understanding of the mechanism behind the pretrained models' improvements, we shift our focus on the analysis of attention weights. We aim to understand directions along which an input molecular representation is decomposed when passed through a given model. In order to do so, we start by considering the Attention Rollout matrix [62] $\tilde{A}$ as a proxy for the model's action on the input (see SI for a more detailed motivation). While this is a strong approximation, it provides a number of non-trivial insights (vide infra). We operate an eigendecomposition of $\tilde{A}$ (from here on we will make use of the bra-ket notation):

$$\tilde{A} = \sum_{i=0}^{N-1} a_i |a_i\rangle\langle a_i| \tag{2}$$

with $a_i \in \mathbb{C}$ and $|a_0| \geq |a_1| \geq ... \geq |a_{N-1}|$ and, based on an empirical observation on one of the pretrained Graphormers (see Fig. 2), we analyze the similarity of the eigenvectors $|a_i\rangle \in \mathbb{C}^n$ with the eigenvectors of the Laplacian matrix L of the input molecular graph decomposed as

$$L = \sum_{i=0}^{N-1} l_i |l_i\rangle\langle l_i| \tag{3}$$

with $l_i \in \mathbb{R}$, $|l_i\rangle \in \mathbb{R}^n$ and $l_0 \leq l_1 \leq ... \leq l_{N-1}$. In particular, by considering the overlap matrix $C_{ij} = |\langle l_i|a_j\rangle|$ we study both how many Laplacian modes are used as models' eigendirections as well as how relevant they are as fraction of the non-trivial spectrum of $\tilde{A}$ (by non-trivial we mean $i \neq 0$ as by construction $|\langle l_0|a_0\rangle| = 1$ for reasons reported in the SI). This fraction is quantified by $\eta = \frac{\sum_{i\in\mathcal{U}\setminus 0}|a_i|}{\sum_{i=1}^{i=N-1}|a_i|}$ where $\mathcal{U} = \{j|\max_j C_{ij} \geq 0.9 \text{ for i} \in (0,1,2,...,\mathrm{N}-1)\}$ with 0.9 being a chosen arbitrary threshold for similarity. Based on these quantities, we define a metric that factors everything together as:

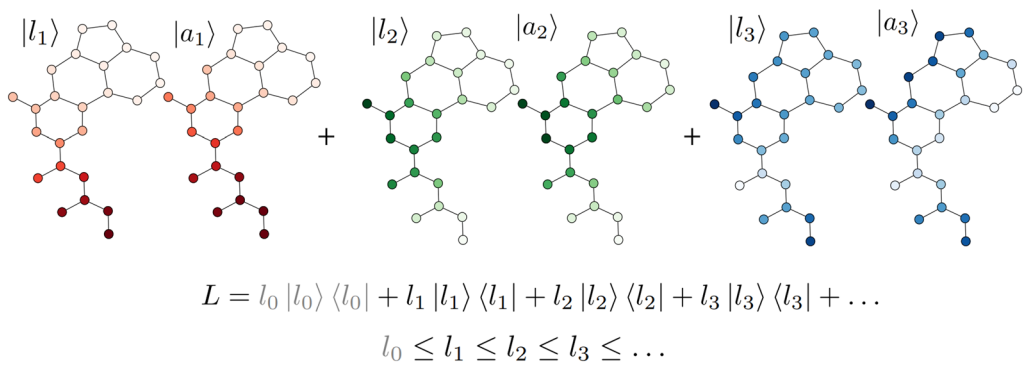

Fig. 2 Visual representation for a molecule in the TDC dataset of the comparison between the most relevant eigenvectors of the Attention Rollout matrix from a model pretrained on atom-level QM properties and the low-frequency eigenvectors of the graph Laplacian associated to the molecular structure. Each eigenvector is a function defined on the graph nodes, and hence the color scale used here corresponds to $|a_i\rangle$ for the eigenvectors of the Attention Rollout matrix $\tilde{A}$ with eigenvalue a_i and to $|l_i\rangle$ for the eigenvectors of the graph Laplacian L with eigenvalue l_i

$$\zeta = \eta \sum_{i=1}^{N-1} \Theta\left(\max_j C_{i,j} - 0.9\right) \tag{4}$$

where Θ is the Heaviside function. We then evaluate ζ averaged over the test set of each downstream task reporting per each group of models the distribution across tasks for fixed pretraining condition. This quantity, as noted in the SI, can be loosely interpreted as an indicator of oversmoothing of the input graph information. A higher ζ, in fact, can be interpreted as a higher bandwidth in the Fourier space defined by the input graph.

Neighbor sensitivity analysis

Following the hypothesis that atomic QM properties provide a good description of the atomic environment around each atom, we carry out a sensitivity analysis to understand in every model how much each atomic representation $\mathrm{GT}(\mathbf{X}) \in \mathbb{R}^{\mathrm{n}\times\mathrm{d}}$ in the last layer is influenced by changes in the input encodings $\mathbf{X} \in \mathbb{R}^{n\times d}$ of the k^{th} order neighbors within the same molecule. These changes are considered in the differential sense using the Jacobian matrix of the respective latent representation. We hence compute the following quantity for 50 randomly selected molecules from the TDC test sets:

$$\mathcal{S}_k = \left\langle \left\langle \sqrt{\sum_{\nu=0}^{d}\sum_{\mu=0}^{d}\left(\frac{\delta \mathrm{GT}(\mathbf{X})_{\mathrm{i},\nu}}{\delta \mathbf{X}_{j,\mu}}\right)^2} \right\rangle_{j\in\mathcal{K}_i} \right\rangle_{i\in\mathcal{M}}, \tag{5}$$

where the μ and ν indices run over the feature dimension d, $\mathcal{K}_i$ is the set of k^{th} order neighbors of the atom i and $\mathcal{M}$ is the set of the atoms in the molecule. As a result, for each molecule, $\mathcal{S}_k$ is the Frobenius norm of the considered Jacobian matrix averaged over all atoms with the chosen topological distance k from the atom i, and averaged over the set of all atoms. A closely related quantity has been studied as indicator of oversquashing in graph message-passing neural networks, a phenomenon where bottlenecks in the message passing mechanism prevent proper information propagation [66]. In our case, it is used as measure of receptive field for the representation of atomic environments obtained with the models, and compared across pretraining methodologies. For each molecule the vector of sensitivities $[\mathcal{S}_0, \mathcal{S}_1, \ldots]$ is then standardized subtracting the minimum value and then dividing by the maximum value (which is usually $\mathcal{S}_0$). These vectors are collected for the sampled structures and the selected models, and the behavior is then analyzed from the first to fifth neighbor atom node.

Results

Benchmark results

Model performances obtained for the downstream tasks are summarized in table 1 where the best model in the tested group is highlighted in bold. We evaluate the best results based on their mean values. Then, for every other model, we perform a t-test paired by seed to test the hypothesis that the best model is significantly better than the others. The models for which the null hypothesis cannot be refuted were highlighted with the only exception being the exclusion of two cases where the standard deviation is one order of magnitude higher than for all the

Table 1 Global results obtained from the ADMET group of TDC

Task	Metric	Scratch	All	Charges	Nmr	Fukui_n	Fukui_e	Masking	Homo-lumo
caco2_wang	MAE↓	0.442 ± 0.041	**0.354 ± 0.015**	**0.404 ± 0.069**	**0.364 ± 0.046**	**0.346 ± 0.034**	0.483 ± 0.036	0.471 ± 0.080	0.381 ± 0.040
hia_hou	ROC-AUC ↑	**0.972 ± 0.015**	**0.982 ± 0.003**	**0.973 ± 0.027**	**0.977 ± 0.011**	0.967 ± 0.011	0.908 ± 0.019	**0.981 ± 0.013**	0.869 ± 0.037
pgp_broccatelli	ROC-AUC ↑	0.892 ± 0.011	**0.913 ± 0.015**	0.902 ± 0.019	**0.917 ± 0.009**	0.896 ± 0.020	**0.911 ± 0.008**	**0.921 ± 0.003**	0.870 ± 0.016
bioavailability_ma	ROC-AUC ↑	0.606 ± 0.040	**0.673 ± 0.028**	**0.662 ± 0.071**	0.640 ± 0.040	0.663 ± 0.025	0.616 ± 0.082	**0.698 ± 0.035**	**0.667 ± 0.031**
lipophilicity_astrazeneca	MAE ↓	0.539 ± 0.036	**0.393 ± 0.005***	**0.425 ± 0.023***	0.424 ± 0.007*	0.457 ± 0.008*	0.463 ± 0.011*	0.462 ± 0.005*	0.451 ± 0.011*
solubility_aqsoldb	MAE ↓	0.878 ± 0.031	**0.720 ± 0.010***	0.726 ± 0.011*	0.728 ± 0.014*	0.756 ± 0.012	0.771 ± 0.015	0.769 ± 0.007	0.772 ± 0.019
bbb_martins	ROC-AUC ↑	0.860 ± 0.016	**0.872 ± 0.021**	**0.874 ± 0.011**	**0.869 ± 0.014**	0.848 ± 0.018	0.845 ± 0.014	**0.861 ± 0.025**	**0.883 ± 0.007**
ppbr_az	MAE ↓	8.477 ± 0.483	**7.589 ± 0.203**	**7.668 ± 0.236**	**7.542 ± 0.215**	**7.530 ± 0.318**	8.026 ± 0.222	**8.056 ± 0.340**	**7.874 ± 0.287**
vdss_lombardo	Spearman ↑	0.554 ± 0.049	0.624 ± 0.020	**0.637 ± 0.022**	0.616 ± 0.034	0.616 ± 0.015	**0.652 ± 0.012**	0.620 ± 0.023	0.580 ± 0.029
cyp2d6_veith	PR-AUC ↑	0.549 ± 0.043	**0.621 ± 0.046**	**0.675 ± 0.014**	**0.643 ± 0.036**	0.660 ± 0.009	0.638 ± 0.011	0.612 ± 0.021	0.612 ± 0.028
cyp3a4_veith	PR-AUC ↑	0.799 ± 0.012	0.797 ± 0.029	**0.847 ± 0.022**	**0.824 ± 0.021**	**0.838 ± 0.016**	**0.828 ± 0.018**	0.817 ± 0.014	0.794 ± 0.018
cyp2c9_veith	PR-AUC ↑	0.706 ± 0.014	0.703 ± 0.022	**0.726 ± 0.024**	**0.739 ± 0.011**	**0.722 ± 0.021**	**0.734 ± 0.014**	**0.736 ± 0.014**	0.708 ± 0.010
cyp2d6_substrate_carbonmangels	PR-AUC ↑	0.546 ± 0.042	**0.648 ± 0.031**	**0.634 ± 0.050**	**0.653 ± 0.023**	**0.619 ± 0.057**	0.578 ± 0.052	**0.677 ± 0.022**	0.582 ± 0.036
cyp3a4_substrate_carbonmangels	ROC-AUC ↑	0.637 ± 0.027	0.630 ± 0.015	0.646 ± 0.020	0.642 ± 0.009	0.645 ± 0.015	0.635 ± 0.031	0.641 ± 0.030	**0.685 ± 0.015**
cyp2c9_substrate_carbonmangels	PR-AUC ↑	0.360 ± 0.022	0.374 ± 0.028	**0.404 ± 0.027**	0.394 ± 0.024	**0.405 ± 0.036**	0.375 ± 0.030	**0.396 ± 0.024**	**0.439 ± 0.043**
half_life_obach	Spearman ↑	0.373 ± 0.076	0.462 ± 0.154	**0.559 ± 0.034**	0.487 ± 0.045	0.486 ± 0.030	0.476 ± 0.015	0.462 ± 0.052	0.426 ± 0.039
clearance_microsome_az	Spearman ↑	0.448 ± 0.038	0.548 ± 0.029	**0.620 ± 0.007**	**0.613 ± 0.014**	0.554 ± 0.019	0.513 ± 0.022	0.555 ± 0.022	0.565 ± 0.032
clearance_hepatocyte_az	Spearman ↑	0.336 ± 0.050	0.382 ± 0.032	**0.456 ± 0.015**	0.460 ± 0.019	0.374 ± 0.021	0.353 ± 0.028	**0.478 ± 0.018**	0.413 ± 0.030
herg	ROC-AUC ↑	0.709 ± 0.080	0.788 ± 0.029	0.824 ± 0.046	0.834 ± 0.030	0.752 ± 0.042	0.758 ± 0.053	**0.880 ± 0.003**	0.790 ± 0.031
ames	ROC-AUC ↑	0.772 ± 0.022	**0.822 ± 0.005**	**0.821 ± 0.010**	**0.833 ± 0.014**	**0.820 ± 0.009**	**0.823 ± 0.012**	0.801 ± 0.008	0.808 ± 0.008
dili	ROC-AUC ↑	0.856 ± 0.037	**0.892 ± 0.033**	**0.859 ± 0.055**	**0.898 ± 0.022**	**0.847 ± 0.016**	0.812 ± 0.122	**0.906 ± 0.021**	0.854 ± 0.017
ld50_zhu	MAE ↓	0.593 ± 0.038	0.559 ± 0.016	0.571 ± 0.012	**0.538 ± 0.014***	0.592 ± 0.029	0.618 ± 0.014	0.577 ± 0.010	0.582 ± 0.031
Number of best models		1	12	17	13	7	5	11	5

Each row corresponds to a specific task, along with the metric used for evaluation. Columns represent different pretrainings considered. Highlighted values denote the best performance achieved among our models, based on the average value and t-tests paired across seeds. Additionally, cases where our results surpass in mean value the top-performing model in the TDC leaderboard are marked with an asterisk (*). For an explanation of model tags we refer to the SI

other results. Overall, while in most cases all the pretraining strategies provide some improvement, pretraining on HLG stands out only for one property, albeit still being among the best models in the group for four more cases. While masking pretraining also significantly outperforms other models only in one case, we find it sharing top performance with other models for ten more downstream tasks. When the models pretrained with atom-level QM properties are considered as a group, we find it to contain the best model overall (at least one better than both masking and HLG) in ten cases, and tying for best model in twenty cases out of twenty-two. Within the group one can see that models pretrained on charges, NMR shifts and all atomic QM properties provide overall a greater number of best results than models pretrained on Fukui functions. Finally, we notice that for the case of solubility, lipophilicity and acute toxicity (LD50) we obtain superior results than the respective best models in the TDC leaderboard.

Table 2 Results of the fine-tuning on internal microsomal clearance dataset

	Metric	Scratch	All	Charges	Nmr	Fukui_n	Fukui_e	Masking	Homo-lumo
Clearance_1	R^2 ↑	0.505 ± 0.010	**0.640 ± 0.004**	0.629 ± 0.006	**0.635 ± 0.006**	0.599 ± 0.004	0.593 ± 0.004	0.580 ± 0.012	0.602 ± 0.006
	Spearman ↑	0.728 ± 0.008	**0.807 ± 0.003**	0.799 ± 0.004	**0.801 ± 0.003**	0.785 ± 0.003	0.785 ± 0.001	0.774 ± 0.007	0.786 ± 0.004
Clearance_2	R^2 ↑	0.534 ± 0.006	**0.653 ± 0.004**	0.633 ± 0.003	**0.643 ± 0.005**	0.598 ± 0.007	0.610 ± 0.008	0.597 ± 0.002	0.607 ± 0.005
	Spearman ↑	0.750 ± 0.005	**0.818 ± 0.003**	0.807 ± 0.004	**0.811 ± 0.002**	0.789 ± 0.002	0.795 ± 0.006	0.786 ± 0.002	0.794 ± 0.002

Results are reported for both values of clearance in the dataset and for all pretraining strategies both in terms of R^2 coefficient and in terms of Spearman's rank coefficient. Highlighted values denote the best performance achieved among our models, based on the average value and t-tests paired across seeds

Performance on internal microsomal clearance data

Although the TDC dataset provides a well established benchmark in modeling ADMET properties, the different models reported here demonstrated close performance on multiple tasks. Expecting divergence of model metrics, we tested our methodology on a much larger dataset of proprietary JNJ HLM clearance data and summarized the results in table 2. The models pretrained on all atomic QM properties obtain the best results in both metrics (R^2 and Spearman's coefficient), followed closely by models pretrained on NMR shifts which are found to not have significantly worse results, and atomic charges. Models pretrained on Fukui indices give the worst results among models pretrained on atomic QM properties, obtaining similar performances to models pretrained on HLG. Notably, and contrary to what seen in the benchmark results, models pretrained using masking obtain the worst results over all pretrained models, albeit still giving improvements over models trained from scratch.

Conservation of pretraining information after fine-tuning

The results obtained on the regularized linear regression of pretraining labels from the representations of the pretraining structures obtained with fine-tuned models are reported in Fig. 3. We report each value of R^2 coefficient with mean and standard deviation over the results obtained from the twenty-two fine-tuned models obtained from each pretraining (masking is excluded from this analysis as it is self-supervised) and also compare them with similar values for non-fine-tuned models. Prior to fine-tuning, the representations exhibit high degree of correlation with other QM properties (Fig. S1), especially those obtained from the model trained on all four atomic QM properties. Despite some degradation, the representations maintain a high degree of linear correlation with their correspondent pretraining property after being fine-tuned on downstream tasks both in absolute terms and with respect to the model trained from scratch. Interestingly, models pretrained on NMR shifts provide highly correlated with the atomic charges representations and also exhibit least degradation of its pretraining information during fine-tuning. On the other hand, models pretrained on Fukui function values and HLG have slightly less linear correlation with the pretraining task, although still being in a quite high range when considering standard deviation and the comparison with the models trained from scratch. Models pretrained on all four atomic properties maintain a high degree of linear correlation with individual properties, but especially with charges and NMR shifts. Finally, we notice the models pretrained on NMR shifts, Fukui electrophilic indices and all atomic properties, exhibit some degree of correlation with HLGs. In contrast, some correlation between the HLG-pretrained representations and atomic properties are observed (Fig. S1) but, after fine-tuning phase, such correlation fully degrade down to levels achieved for a scratch model. We hypothesize that the way the network is trained might be responsible for this asymmetric behavior. Training on atomic properties followed by fine-tuning on downstream molecular properties impacts, in the final layer, both the latent representations relative to the atoms and the one relative to the CLS token used to model molecular properties. This is not the case for HLGs pretrained models, as, in the final layer, only the representation associated with CLS tokens are directly used to compute the output value.

Latent expressivity as across layers

The results of the analysis of ρ_L across layers are summarized for the models using the three main pretraining strategies (all atom-level quantum properties, HLG and masking) and for the models trained from scratch in Fig. 4, while a similar plot comparing the models pretrained on each atom-level QM property separately is reported in the SI. It is evident from the plot that the trend of ρ_L is very different across pretraining methods. Overall, when comparing to the models trained from scratch, all pretraining strategies mitigate the collapse in latent expressivity. In particular, while models pretrained on HLG are characterized by a constant level of expressivity across layers, models pretrained using masking have more similar atomic latent representations in the first few layers and more dissimilar in the last ones. For models pretrained on all atomic quantum properties, a strong increase in expressivity is observed in the first part of the network, reaching a higher value of ρ_L

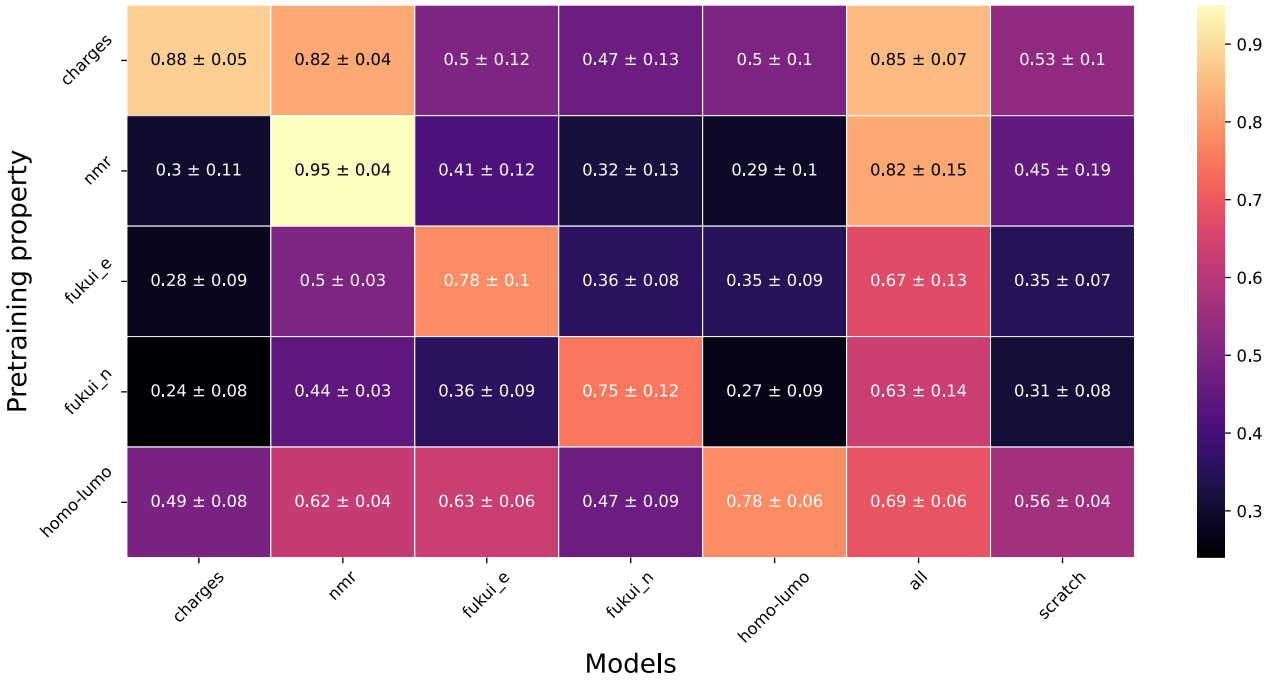

Fig. 3 R^2 for the regression tasks using the representations of a sample of the pretraining data obtained with fine-tuned models. We report the mean and standard deviation over all fine-tuning cases (mean and standard deviation over twenty-two cases)

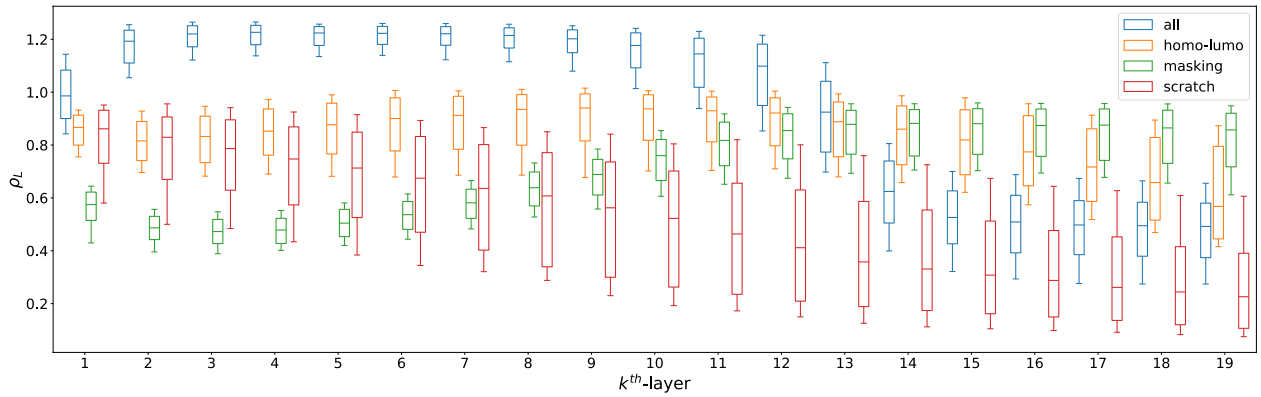

Fig. 4 Expressivity of the latent representation measured with the quantity ρ_L as a function of layer number. This quantity is computed for a sample of 2200 structures extracted uniformly from all the fine-tuning test sets (100 structures for each of the 22 tasks) and results are reported as boxplots at each layer. This is done for models pretrained on HLG, models pretrained on all atom-level QM properties, models pretrained with masking and models trained from scratch. The whiskers go from the 15th percentile to the 85th for better visualization of trends and outliers are excluded for the same reason

than the other cases, followed by a decrease in the last part, closer to the regression head. Regarding pretraining on individual atomic properties, we find that the NMR shifts and charges models have similar behavior. The ones pretrained on Fukui indices, while presenting a similar trend, achieve a lower maximum expressivity more similar to the models pretrained on HLG. The absence of complete expressivity collapse for all pretrained models likely comes from the much higher number of examples that the models were trained on comparably to the models trained from scratch. However, dissimilar behaviors of expressivity indicate that the pretraining strategies explored here produce very different models even when the overall performance improvements on the benchmark are comparable. It is notable that if we consider the highest value across layers for each model, models pretrained on atomic quantum properties achieve the highest maximum latent expressivity. While we do not have a

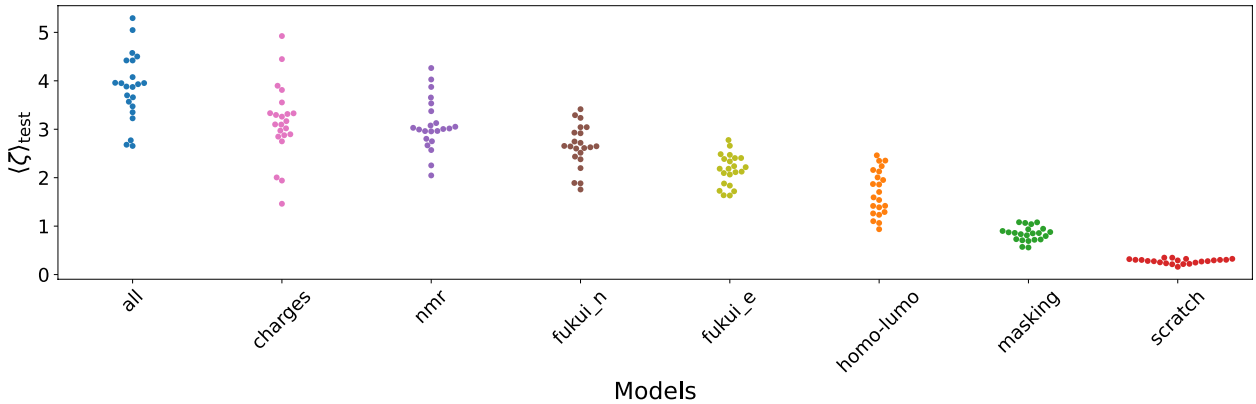

Fig. 5 Spectral perception of the input graphs for the models fine-tuned on the TDC datasets grouped by pretraining strategy. This is reported in the form of swarm plots of the values of ζ averaged across each of the 22 fine-tuning test sets for fixed pretraining strategy

definitive explanation for the final sharp decrease in the last layers, we hypothesize that quantum atomic property regression requires the model to capture correlations between atoms within the same molecular structure when close to the last layer, as these properties depend on the surrounding atoms as do their respective properties. This would be sustained from the seemingly opposite trend found in the models pretrained with masking, which is a classification task that requires to maximally distinguish atomic latent representations close to the last network layer using a cross entropy loss function that rewards higher certainties.

Spectral analysis of attention rollout

We evaluate the metric ζ defined in Eq. 4 as described in the Methods section obtaining a distribution of 22 values over the downstream tasks per each group of studied models. The results are reported in Fig. 5 as a set of swarm plots. Firstly, it is evident that the models trained from scratch present values of ζ that are close to 0 indicating little to no presence of non-trivial Laplacian eigenmodes in the spectrum of their $\tilde{A}$ matrix. On the contrary, every pretrained model (including masking) presents nonzero values of ζ across the downstream tasks raging from ~ 1 to ~ 6. Within this last group of models one can clearly notice how pretraining on the atom-level QM properties provides the strongest increase of perception of graph Laplacian eigenmodes. In particular, the model pretrained using all properties in a multi-task fashion presents the highest values of ζ, followed by the models pretrained on charges, NMR shifts, nucleophilic and electrophilic Fukui function indices. The models pretrained on HLGs also present some degree of spectral perception, albeit in a lower range than the previously mentioned models, followed by models pretrained using masking which present the lowest graph spectral perception among the set of pretrained models.

Neighbor sensitivity analysis

The results of the neighbor sensitivity analysis are reported in Fig. 6. For each considered group of models we report the value of $\mathcal{S}_k$ for $k \in [1, \ldots, 5]$ in boxplots over 1100 structures sampled uniformly from the test sets of the fine-tuning tasks. It is found that the models trained from scratch exhibits a constant and low sensitivity of representation with respect to neighboring atoms, whereas pretrained models present a reasonable descending trend with topological distance. In particular, the models pretrained on all the atomic QM properties have a stronger sensitivity than all other models for all the considered topological distances, especially, for first and second neighbors. The models pretrained on HLG present slightly higher sensitivities than the ones pretrained using masking which presents the lowest set of sensitivities among all pretraining strategies. Among the models trained on individual atomic QM properties, the sensitivity ranges tend to overlap, but they are positioned in between the models pretrained on HLGs and models pretrained on all atomic QM properties for all considered topological distances.

Discussion

Our in-depth analysis demonstrates that, among the tested strategies, pretraining the Graphormer on four atomic QM properties in the multitask fashion provides the best model for subsequent fine-tuning on ADMET properties. The final models exhibit high performance results in the TDC benchmark and, more importantly, outperform other models on a much larger dataset of JNJ HLM clearance data. The latent space analysis also

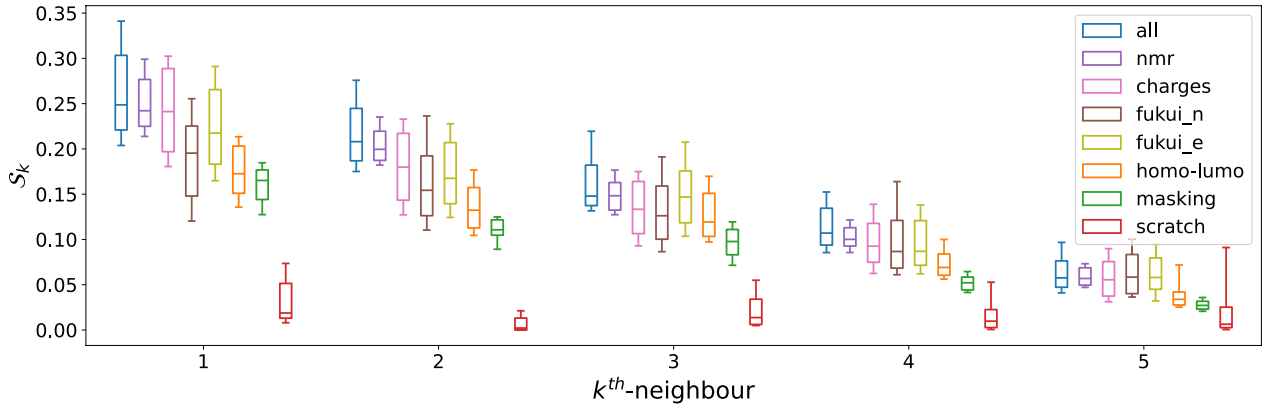

Fig. 6 Boxplots of the k^{th} neighbor normalized sensitivities $\mathcal{S}_k$ for $k \in [1, \ldots, 5]$. Each boxplot summarizes a sample of 1100 structures extracted uniformly from all the fine-tuning test sets (50 structures for each of the 22 tasks). We report this quantity for all studied pretraining strategies, and also for the models trained from scratch. The whiskers cover the values from the 15th percentile to the 85th for better visualization of trends. Outliers are excluded for the same reason

positions the respective models at the top with highest values of latent expressivity, neighbor sensitivities, and graph-spectral perception. Besides, last layer representations of the respective models pretrained on four atomic properties retain high degree of correlation with all types of pretraining atomic data after fine-tuning.

Pretraining on NMR shielding constants and atomic charges yields the models that are sharing the second overall rank in studied metrics. Moreover, these pretrainings also provide the highest number of top-performing models in the TDC dataset. Interestingly, while pretraining charge labels can be modeled with good results using NMR-pretrained models after downstream tasks fine-tuning, the opposite doesn't hold, which indicates that NMR shifts may contain richer information than atomic charges. NMR chemical shifts are indeed known for their extreme sensitivity to the atomic environment of the respective nuclei, covering both electronic and to some extent steric effects.

The Fukui indices pretraining posses the third cumulative rank among the studied approaches. Unexpectedly, the respective models were not among the top performing models for clearance modeling, despite strong relevance of electrophilic Fukui indices to site of metabolism predictions[34] which in turn drives the hepatic clearance of drugs. Fukui indices are often calculated with the aim to build affinity QSAR models, particularly to CYP enzymes [77]. This observation suggests that there is more to pretraining such complex models than only transfer learning between tasks.

Models pretrained on HLG resulted being close to models pretrained on Fukui indices under the performance indicators on both TDC and JNJ HLM data, and just slightly worse under representation indicators. While the relevance of HLG to ADMET properties is arguable, it was used as a feature for QSAR modeling of CYP enzyme inhibition[78, 79] and generally characterizes the propensity of a molecule to donate or accept electrons, sometimes referred as a global hardness [80]. In particular, Fukui indices and frontier orbital levels are tightly related properties and often calculated to characterize biologically active molecules. When comparing HLG pretraining to Fukui indices pretraining, though, it is important to notice that the atomic QM properties dataset contains $\sim$ 20 times less molecular structures than the HLG pretraining dataset, making atomic properties much more efficient in terms of training time and resources. At the same time, because each molecule in the atomic properties pretraining dataset contains on average 17 non-hydrogen atoms per molecule, the overall number of data for each atomic property is on the same scale as the HLG dataset. Furthermore, because calculation of atomic properties typically requires only a fraction of overall computational resources spent on geometry optimization and electronic structure refinement during QM modeling, such properties provide a finer grade physical description of molecular structures with a non-dramatic overhead in the data generation phase. Little or no overlap (as shown in Fig. 1) between the chemical spaces of pretraining and downstream task datasets, does not block performance improvements obtained from pretraining. We speculate that pretraining on atom-resolved QM properties incorporates fine-grade knowledge about the electronic structure of molecules that better generalizes to molecules outside of pretraining domains and results in apparent higher efficiency in downstream tasks.

Masking pretraining, considered as example of label-free atom-level pretraining, provided inconsistent results. If the performance on the TDC benchmark is comparable with other models pretrained on atomic QM properties, the analysis of the latent representations together with the results on the much wider JNJ HLM dataset place this pretraining strategy at the bottom among the tested ones, confirming that the improvement seen with atom-level QM pretraining does not come solely from being atom-resolved. It is worth noting that for consistency we utilized the same set of molecules for the masking pretraining as for the pretraining on atomic QM properties. Considering simplicity of data preparation, masking pretraining can be used with much larger datasets containing tens of millions of molecules and potentially improve the performance of the respective models, however such experiments were beyond the scope of the present study.

These last findings furthermore highlight the limitation of picking the best pretraining method solely using the results obtained on public benchmark datasets and the importance of utilizing other metrics. In this regard, we would like to emphasize that novel analysis of the spectrum of the Attention Rollout matrix documented a non-trivial effect arising in pretrained GTs where the model, albeit under the strong approximation of the explainability method, shows hints of filtered spectral graph convolution. Such findings connect the GT architecture to the family of Spectral Graph Neural Networks (SGNNs) [81–84], and should stimulate further research potentially leading to the development of more robust models that can better leverage both on the graph-spectral features typical of SGNNs and the flexibility of transformer-based architectures for graph-based applications.

Conclusions

In this work we explored the effects of pretraining deep Graph Transformer models on quantum chemical data for improving the performance of modeling ADMET properties of drug-like molecules as downstream tasks. Atomic properties such as charges and a molecular property, HOMO-LUMO gap, were evaluated along with atom masking pretraining, an analogue of a well know self-supervised method for language models. Pretrained models almost always showed much better accuracy than the scratch models. The results on the public ADMET benchmark show that, in general, an atom-resolved pretraining, both on QM properties and via masking, obtains better performances than pretraining on a HOMO-LUMO gap. Comparable performance of the atom masking pretraining is not confirmed when moving to a larger proprietary human liver microsome intrinsic clearance dataset, revealing pretraining on atomic charges and NMR shifts as the most superior techniques and highlighting that relative performances on relatively small public benchmarks may not hold when scaled up. Multiple analyses were then performed on the models' weights, revealing several insights into the inner workings of the differently pretrained models. Firstly, it is found that pretraining information is, in general, preserved after fine tuning albeit to different degrees depending on the pretraining. Secondly, the well-known phenomenon of rank collapse in the latent representation of deep models is hindered in different ways by different pretraining strategies, showing that even models with similar performances can have very different inner mechanisms of feature extraction. Thirdly, analyzing the spectral properties of the attention weights using a custom metric, we find that models pretrained on atom-level QM properties can capture more low-frequency Laplacian eigenmodes of the input molecular graphs in what seems to be a low-pass filtering behavior. Finally, by studying the Jacobian of the model from the first to the last layer and devising a specific metric, we show how models pretrained with atom-resolved QM properties achieve better representations of the chemical environments around the atoms by showing that atomic representations have a higher sensitivity with respect to the neighboring atoms. As a global observation, when modeling public TDC ADMET data, the latent representation analyses rather than model performance metrics give results that are more aligned with the performances on the larger internal HLM intrinsic clearance dataset. While we were unable to provide explanations for all the collected observations, we hope that the present work provides a different perspective on evaluating molecular property modeling, as well as valuable insights for future research in molecular representation learning and for the development of useful in-silico datasets.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00970-0.

Supplementary material 1.

Acknowledgements

We thank Dr. Leonardo Medrano Sandonas for the helpful comments. A.F. thanks Dr. Matthieu Sarkis for the discussions around mathematical properties of graph Laplacians.

Author contributions

A.F. wrote the code for the paper results, conducted model training and evaluation, devised and performed all representation analyses. R.N. conceived the idea of atom-level quantum mechanical properties pretraining, developed the chytorch version of Graphormer, curated datasets, and provided code support and deep learning advice. J.A.-M. proposed the use of the TDC benchmark, contributed to preliminary pretraining and fine-tuning results,

offered deep learning and code guidance. K.C. contributed to the Attention Rollout spectral analysis and supervised all stages of the project. All authors actively discussed the results and contributed to the final manuscript.

Funding
This research was financially supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956832, "Advanced Machine learning for Innovative Drug Discovery" (AIDD).

Availability of data and materials
The code pertaining the results on public data will be made available on https://github.com/aidd-msca/GraphQPT.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg , Luxembourg. [2]Drug Discovery Data Sciences, Janssen Pharmaceutica NV, Turnhoutseweg 30, 2340 Beerse, Belgium. [3]Johnson & Johnson Innovative Medicine, 301 Binney Street, Cambridge, MA 02142, USA.

Received: 11 October 2024 Accepted: 7 February 2025
Published online: 27 February 2025

References

1. Rupp M, Tkatchenko A, Müller K-R, von Lilienfeld OA (2012) Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett 108:058301
2. Hansen K, Biegler F, Ramakrishnan R, Pronobis W, von Lilienfeld OA, Müller K-R, Tkatchenko A (2015) Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. J Phys Chem Lett 6:2326–2331
3. Huang B, Symonds NO, von Lilienfeld OA (2020) Handbook of materials modeling: methods: theory and modeling. Springer International Publishing, Cham, pp 1883–1909
4. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug Discov Today 23:1241–1250
5. Jayatunga MK, Xie W, Ruder L, Schulze U, Meier C (2022) AI in small-molecule drug discovery: a coming wave. Nat Rev Drug Discov 21:175–176
6. Bule M, Jalalimanesh N, Bayrami Z, Baeeri M, Abdollahi M (2021) The rise of deep learning and transformations in bioactivity prediction power of molecular modeling tools. Chem Biol Drug Des 98:954–967
7. Li MM, Huang K, Zitnik M (2022) Graph representation learning in biomedicine and healthcare. Nat Biomed Eng 6:1353–1369
8. Chuang KV, Gunsalus LM, Keiser MJ (2020) Learning molecular representations for medicinal chemistry. J Med Chem 63:8705–8722
9. Born J, Markert G, Janakarajan N, Kimber TB, Volkamer A, Martínez MR, Manica M (2023) Chemical representation learning for toxicity prediction. Digit Discov 2:674–691
10. Wang Y, Wang J, Cao Z, Barati Farimani A (2022) Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell 4:279–287
11. Kaufman B, Williams EC, Underkoffler C, Pederson R, Mardirossian N, Watson I, Parkhill J (2024) COATI: multimodal contrastive pretraining for representing and traversing chemical space. J Chem Inf Model 64:1145–1157
12. Ilnicka A, Schneider G (2023) Compression of molecular fingerprints with autoencoder networks. Mol Inf 42:2300059
13. Sanchez-Fernandez A, Rumetshofer E, Hochreiter S, Klambauer G. Contrastive learning of image- and structure-based representations in drug discovery. ICLR2022 Machine Learning for Drug Discovery. 2022
14. Fang Y, Zhang Q, Zhang N, Chen Z, Zhuang X, Shao X, Fan X, Chen H (2023) Knowledge graph-enhanced molecular contrastive learning with functional prompt. Nat Mach Intell 5:542–553
15. Wen J et al (2023) Multimodal representation learning for predicting molecule-disease relations. Bioinformatics 39:085
16. Su B, Du D, Yang Z, Zhou Y, Li J, Rao A, Sun H, Lu Z, Wen J-R. A molecular multimodal foundation model associating molecule graphs with natural language. 2022. arXiv:2209.05481
17. Wang Z, Jiang T, Wang J, Xuan Q. Multi-modal representation learning for molecular property prediction: sequence, graph, geometry. 2024. arXiv:2401.03369
18. Gao X, Ramezanghorbani F, Isayev O, Smith JS, Roitberg AE (2020) TorchANI: a free and open source PyTorch-based deep learning implementation of the ANI neural network potentials. J Chem Inf Model 60:3408–3415
19. Schütt K, Kindermans P-J, Sauceda Felix HE, Chmiela S, Tkatchenko A, Müller K-R (2017) SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. Adv Neural Inf Process Syst 30:992–1002
20. Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, Molinari N, Smidt TE, Kozinsky B (2022) E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nat Commun 13:2453
21. Deng J, Yang Z, HW, et al (2023) A systematic study of key elements underlying molecular property prediction. Nat Commun 14:6395
22. Dou B, Zhu Z, Merkurjev E, Ke L, Chen L, Jiang J, Zhu Y, Liu J, Zhang B, Wei G-W (2023) Machine learning methods for small data challenges in molecular science. Chem Rev 123:8736–8780
23. Glavatskikh M, Leguy J, Hunault G, Cauchy T, Da Mota B (2019) Dataset's chemical diversity limits the generalizability of machine learning predictions. J Cheminform 11:69
24. Ektefaie Y, Shen A, Bykova D, Marin M, Zitnik M, Farhat M. Evaluating generalizability of artificial intelligence models for molecular datasets. 2024. *bioRxiv*
25. Broccatelli F, Trager R, Reutlinger M, Karypis G, Li M (2022) Benchmarking accuracy and generalizability of four graph neural networks using large in vitro ADME datasets from different chemical spaces. Mol Inf 41:2100321
26. David Z, Huang JCB, Bahmanyar SS (2021) The challenges of generalizability in artificial intelligence for ADME/Tox endpoint and activity prediction. Expert Opin Drug Discov 16:1045–1056
27. Keith JA, Vassilev-Galindo V, Cheng B, Chmiela S, Gastegger M, Müller K-R, Tkatchenko A (2021) Combining machine learning and computational chemistry for predictive insights into chemical systems. Chem Rev 121:9816–9872
28. OpenAI et al. GPT-4 Technical Report. 2024. arXiv:2303.08774
29. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: Open and Efficient Foundation Language Models. 2023. arXiv:2302.13971
30. Wang Y, Xu C, Li Z, Barati Farimani A (2023) Denoise pretraining on nonequilibrium molecules for accurate and transferable neural potentials. J Chem Theory Comput 19:5077–5087
31. Xia J, Zhao C, Hu B, Gao Z, Tan C, Liu Y, Li S, Li SZ. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules. The Eleventh International Conference on Learning Representations. 2023
32. Xia J, Zhu Y, Du Y, Li SZ. A Systematic Survey of Chemical Pre-trained Models. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23. Survey Track. 2023. pp 6787–6795
33. Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, Leskovec J. Strategies for Pre-training Graph Neural Networks. International Conference on Learning Representations. 2020
34. Beck ME (2005) Do fukui function maxima relate to sites of metabolism? A critical case study. J Chem Inf Model 45:273–282
35. Wang X, Wang L, Wang S, Ren Y, Chen W, Li X, Han P, Song T (2023) QuantumTox: utilizing quantum chemistry with ensemble learning for molecular toxicity prediction. Comput Biol Med 157:106744

36. Beck ME, Schindler M (2007) Pesticide Chemistry. John Wiley & Sons, Ltd, Hoboken, pp 227–238
37. Göller AH (2019) The art of atom descriptor design. Drug Discov Today Technol 32–33:37–43
38. Hoja J, Medrano Sandonas L, Ernst BG, Vazquez-Mayagoitia A, DiStasio RA Jr, Tkatchenko A (2021) QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. Sci Data 8:43
39. Medrano Sandonas L, Van Rompaey D, Fallani A, Hilfiker M, Hahn D, Perez-Benito L, Verhoeven J, Tresadern G, Kurt Wegner J, Ceulemans H, Tkatchenko A (2024) Dataset for quantum-mechanical exploration of conformers and solvent effects in large drug-like molecules. Sci Data 11:742
40. Isert C, Atz K, Jiménez-Luna J, Schneider G, QMugs, (2022) quantum mechanical properties of drug-like molecules. Sci Data 9:273
41. Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, Rühl S, Wolverton C (2015) The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. NPJ Comput Mater 1:15010
42. Nakata M, Shimazaki T, Hashimoto M, Maeda T (2020) PubChemQC PM6: data sets of 221 million molecules with optimized molecular geometries and electronic properties. J Chem Inf Model 60:5891–5899
43. Chanussot L et al (2021) Open catalyst 2020 (OC20) dataset and community challenges. ACS Catal 11:6059–6072
44. Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller K-R (2017) Machine learning of accurate energy-conserving molecular force fields. Sci Adv 3:e1603015
45. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera RS, Gold-Parker A, Vogt L, Brockway AM, Aspuru-Guzik A (2011) The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J Phys Chem Lett 2:2241–2251
46. Mobley DL, Guthrie JP (2014) FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput Aided Mol Des 28:711–720
47. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J (2015) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 44:D1045–D1053
48. Smith JS, Nebgen B, Lubbers N, Isayev O, Roitberg AE (2018) Less is more: sampling chemical space with active learning. J Chem Phys 148:241733
49. Smith JS, Isayev O, Roitberg AE (2017) ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. Sci Data 4:170193
50. Kläser K, Banaszewski B, Maddrell-Mander S, McLean C, Müller L, Parviz A, Huang S, Fitzgibbon A. `MiniMol`: A Parameter-Efficient Foundation Model for Molecular Learning. 2024. arXiv:2404.14986
51. Kim J, Chang W, Ji H, Joung I (2024) Quantum-informed molecular representation learning enhancing ADMET property prediction. J Chem Inf Model 64:5028–5040
52. Raja A, Zhao H, Tyrchan C, Nittinger E, Bronstein MM, Deane C, Morris GM. On the Effectiveness of Quantum Chemistry Pre-training for Pharmacological Property Prediction. ICML 2024 AI for Science Workshop. 2024
53. Beaini D et al. Towards foundational models for molecular learning on large-scale multi-task datasets. The Twelfth International Conference on Learning Representations. 2024
54. Lim MA, Yang S, Mai H, Cheng AC (2022) Exploring deep learning of quantum chemical properties for absorption, distribution, metabolism, and excretion predictions. J Chem Inf Model 62:6336–6341
55. Shoghi N, Kolluru A, Kitchin JR, Ulissi ZW, Zitnick CL, Wood BM. From molecules to materials: pre-training large generalizable models for atomic property prediction. ICLR. 2024
56. Ying C, Cai T, Luo S, Zheng S, Ke G, He D, Shen Y, Liu T-Y (2021) Do transformers really perform badly for graph representation? Adv Neural Inf Process Syst 34:28877–28888
57. Nugmanov R, Dyubankova N, Gedich A, Wegner JK (2022) Bidirectional graphormer for reactivity understanding: neural network trained to reaction atom-to-atom mapping task. J Chem Inf Model 62:3307–3315
58. Müller L, Galkin M, Morris C, Rampášek L. Attending to Graph Transformers. *Transactions on Machine Learning Research*. 2024
59. Guan Y, Coley CW, Wu H, Ranasinghe D, Heid E, Struble TJ, Pattanaik L, Green WH, Jensen KF (2021) Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. Chem Sci 12:2198–2208
60. Nakata M, Shimazaki T (2017) PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry. J Chem Inf Model 57:1300–1308
61. Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, Coley CW, Xiao C, Sun J, Zitnik M (2022) Artificial intelligence foundation for therapeutic science. Nat Chem Biol 18:1033–1036
62. Abnar S, Zuidema WH. Quantifying Attention Flow in Transformers. 2020. arXiv:2005.00928
63. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018
64. Fabian B, Edlich T, Gaspar H, Segler MHS, Meyers J, Fiscato M, Ahmed M. Molecular representation learning with language models and domain-relevant auxiliary tasks. CoRR. 2020. arXiv:2011.13230
65. Dong Y, Cordonnier J-B, Loukas A. Attention is Not All You Need: Pure Attention Loses Rank Doubly Exponentially with Depth. 2023. arXiv:2103.03404
66. Topping J, Giovanni FD, Chamberlain BP, Dong X, Bronstein MM. Understanding over-squashing and bottlenecks on graphs via curvature. International Conference on Learning Representations. 2022
67. Arjona-Medina J, Nugmanov R. Analysis of Atom-level pretraining with Quantum Mechanics (QM) data for Graph Neural Networks Molecular property models. 2024. arXiv:2405.14837
68. Fallani A, Arjona-Medina J, Chernichenko K, Nugmanov R, Wegner JK, Tkatchenko A. Atom-Level Quantum Pretraining Enhances the Spectral Perception of Molecular Graphs in Graphormer. AI in Drug Discovery. Cham, 2025. pp 71–81
69. Fabian B, Edlich T, Gaspar H, Segler M, Meyers J, Fiscato M, Ahmed M (2020) Molecular representation learning with language models and domain-relevant auxiliary tasks. Proc. NeurIPS 2020 Workshop on Machine Learning for Molecules. 2020
70. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. arXiv:1802.03426
71. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 67:301–320
72. Shi H, GAO J, Xu H, Liang X, Li Z, Kong L, Lee SMS, Kwok J. Revisiting Over-smoothing in BERT from the Perspective of Graph. International Conference on Learning Representations. 2022
73. Noci L, Li C, Li M, He B, Hofmann T, Maddison CJ, Roy D (2023) The shaped transformer: attention models in the infinite depth-and-width limit. Adv Neural Inf Process Syst 36:54250–54281
74. Noci L, Anagnostidis S, Biggio L, Orvieto A, Singh SP, Lucchi A (2022) Signal propagation in transformers: theoretical perspectives and the role of rank collapse. Adv Neural Inf Process Syst 35:27198–27211
75. Roth A, Bause F, Kriege NM, Liebig T. Preventing Representational Rank Collapse in MPNNs by Splitting the Computational Graph. The Third Learning on Graphs Conference. 2024
76. Roth A, Liebig T. Rank Collapse Causes Over-Smoothing and Over-Correlation in Graph Neural Networks. Proceedings of the Second Learning on Graphs Conference. 2024; pp 35:1–35:23
77. Van Damme S, Bultinck P (2009) Conceptual DFT properties-based 3D QSAR: analysis of inhibitors of the nicotine metabolizing CYP2A6 enzyme. J Comput Chem 30:1749–1757
78. Lewis DF (1997) Quantitative structure-activity relationships in substrates, inducers, and inhibitors of cytochrome P4501 (CYP1). Drug Metab Rev 29:589–650
79. Ai C, Li Y, Wang Y, Li W, Dong P, Ge G, Yang L (2010) Investigation of binding features: effects on the interaction between CYP2A6 and inhibitors. J Comput Chem 31:1822–1831
80. Kaya S, Kaya C (2015) A new method for calculation of molecular hardness: a theoretical study. Comput Theor Chem 1060:66–70
81. Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral Networks and Locally Connected Networks on Graphs. 2014. arXiv:1312.6203
82. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. 2017. arXiv:1609.02907

83. Hammond DK, Vandergheynst P, Gribonval R. Wavelets on Graphs via Spectral Graph Theory. 2009. arXiv:0912.3848
84. Defferrard M, Bresson X, Vandergheynst P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. 2017. arXiv:1606.09375

Publisher's Note