https://doi.org/10.1038/s42004-024-01169-4

Fast and effective molecular property prediction with transferability map

Check for updates

Shaolun Yao[1,2,3], Jie Song[3,4], Lingxiang Jia[2], Lechao Cheng[5], Zipeng Zhong[2], Mingli Song[2,3] & Zunlei Feng[3,4]

Effective transfer learning for molecular property prediction has shown considerable strength in addressing insufficient labeled molecules. Many existing methods either disregard the quantitative relationship between source and target properties, risking negative transfer, or require intensive training on target tasks. To quantify transferability concerning task-relatedness, we propose Principal Gradient-based Measurement (PGM) for transferring molecular property prediction ability. First, we design an optimization-free scheme to calculate a principal gradient for approximating the direction of model optimization on a molecular property prediction dataset. We have analyzed the close connection between the principal gradient and model optimization through mathematical proof. PGM measures the transferability as the distance between the principal gradient obtained from the source dataset and that derived from the target dataset. Then, we perform PGM on various molecular property prediction datasets to build a quantitative transferability map for source dataset selection. Finally, we evaluate PGM on multiple combinations of transfer learning tasks across 12 benchmark molecular property prediction datasets and demonstrate that it can serve as fast and effective guidance to improve the performance of a target task. This work contributes to more efficient discovery of drugs, materials, and catalysts by offering a task-relatedness quantification prior to transfer learning and understanding the relationship between chemical properties.

Molecular property prediction, which involves identifying molecules with desired properties[1,2], poses a critical challenge prevalent across various scientific fields. It holds particular significance in chemistry for designing drugs, catalysts, and materials. In recent years, artificial intelligence (AI) technologies have come mainstream in this area, and AI-guided chemical design can efficiently explore chemical space while improving performance based on experimental feedback, showing promise from laboratory research to real-world industry applications[3]. However, it is common that the experimental data size is small as producing labeled data requires time-consuming and expensive experiments[4,5]. In contrast, transfer learning[6] has become a powerful paradigm for addressing data scarcity problem by exploiting the knowledge from related datasets across fields such as natural language processing[7,8], computer vision[9,10], and biomedcine[11,12]. In chemistry, transfer learning leverages pre-trained models on extensive or related datasets to facilitate efficient exploration of vast chemical space[13,14] for various downstream tasks. It has been used to predict properties[15,16], plan synthesis[17,18], and explore the space of chemical reactions[19–22].

Transfer learning can enhance molecular property prediction in limited data sets by borrowing knowledge from sufficient source data sets, thus improving both model accuracy and computation efficiency. Although several previous works have explored the power of transfer learning to enhance molecular property prediction[11,12,23–25], challenges remain. One major challenge is negative transfer, which occurs when the performance after transfer learning is adversely affected due to minimal similarity between the source and target tasks[26,27]. For example, Hu et al.[23] observed that pretrained GNN (at both node-level and graph-level) performed well but yielded negative transfer when pretrained at the level of either entire graphs or individual nodes. Additionally, some supervised pre-training tasks unrelated to the downstream task of interest can even degrade the downstream performance[23,28].

[1]Collaborative Innovation Center of Artificial Intelligence by MOE and Zhejiang Provincial Government, Zhejiang University, 310027 Hangzhou, China. [2]College of Computer Science and Technology, Zhejiang University, 310027 Hangzhou, China. [3]Shanghai Institute for Advanced Study of Zhejiang University, 201203 Shanghai, China. [4]School of Software Technology, Zhejiang University, 315048 Ningbo, China. [5]School of Computer Science and Information Engineering, Hefei University of Technology, 230009 Hefei, China. e-mail: zunleifeng@zju.edu.cn

Negative transfer primarily stems from suboptimal model and layer choices, as well as insufficient task relatedness, highlighting the need to evaluate transferability prior to applying transfer learning. In computer vision, some researchers have recently focused on selecting the best model from a pool of options by estimating the transferability of each model[29–32]. In molecular property prediction, recent efforts involve investigating the relatedness of the source task to the target task. To maximize the performance on a target task and prevent negative transfer, existing methods mainly rely on a molecular distance metric to measure the similarity of molecules, such as Tanimoto coefficient (based on molecular fingerprint)[33,34] and a chemical distance measure (based on fingerprint and subgraph)[35]. Moreover, inspired by the seminal work of Taskonomy[36] and subsequent works such as RSA[37], an emerging approach in modeling the similarity between computer vision tasks, there have been some recent attempts towards developing representation similarity measurement to quantify similarity between biological compounds, such as molecules[38,39], proteins[40], and macromolecules[41]. These techniques leverage pre-trained model representations to assess biological compound similarity across different tasks. As they stand in need of model optimization across all datasets, and their computational costs that are as high as fine-tuning with target tasks exclude their applications in quantifying transferability prior to fine-tuning. In addition, there is still no metric for quantifying how suitable the source property is for the target property prior to training on the target task in a computation-efficient manner.

To this end, we put forward a simple, fast, and effective Principal Gradient-based Measurement (PGM) to quantify the transferability from the source property to the target property (Fig. 1). First, to approximate the direction of model optimization on a molecular property prediction dataset, we design a restart scheme to calculate a principal gradient in an optimization-free manner. The distance between the principal gradient obtained from model training on the source dataset and that derived from the target dataset indicates transferability. Second, we build a quantitative transferability map by performing PGM on various molecular property prediction datasets to show the inter-property correlations in property space distribution. Third, through the map, we can capture and transfer the most desirable source dataset for the given target dataset, so as to promote performance on the target task and avoid negative transfer. To verify the effectiveness of the proposed PGM, we evaluated PGM thoroughly on 12 benchmark datasets from MoleculeNet[42] with various molecular property prediction tasks. Comprehensive experiments on multiple combinations of transfer learning tasks demonstrate that the quantitative transferability derived from PGM is strongly related to the transfer learning performance. The proposed approach can serve as fast and effective guidance to enhance the transfer performance of molecular property prediction. Our contributions can be summarized as follows:

- We propose a molecular property transferability measurement technique, termed as PGM, which can rapidly and effectively measure the transferability between the source and target molecular property prediction datatsets.
- Furthermore, we perform PGM on molecular benchmarks to build a transferability map that quantifies inter-property correlations. The map is extensible and can be a reference standard for transfer learning in molecular property prediction, even when applied to a few target samples.
- We empirically demonstrate that the transferability measured by the proposed technique has a strong correlation with the transfer learning performance of molecular property prediction tasks in reality.

Significantly, the proposed PGM has another two advantages: (1) it is computation-efficient, free of either calculating by brute force or model optimization. (2) it includes model-agnostic molecular encoder and predictor, thereby applied to various machine learning model frameworks.

Results

Overview of PGM guided transfer learning

In this paper, we introduce Principal Gradient-based Measurement (PGM) to quantify transferability between the source and target molecular properties. The framework of PGM guided transfer learning comprises three components as illustrated in Fig. 1: (1) PGM, (2) transferability map building, and (3) application of transfer learning.

PGM. The model optimization process aims to move from the initial point towards the optimal convergence point (Fig. 1a), where the model's parameters minimize the loss function and the model achieves peak performance. This optimization process involves iteratively updating the model parameters to reduce the loss function, commonly employing gradient descent as the preferred method. Building upon this conceptualization, the model is initialized from the same initial point and subsequently optimized on both source and target molecular property prediction datasets. The distance between the optimal convergence point of the source task and that of the target task refers to transferability. A smaller distance (between source property A and the target property, colored in orange and red, respectively) suggests a higher level of task similarity. Based on the invisible optimal convergence point, computing the transferability by optimizing models is as prohibitively expensive as training models on the given target dataset and all alternative source datasets, while the transferability offers benefits only when it can be

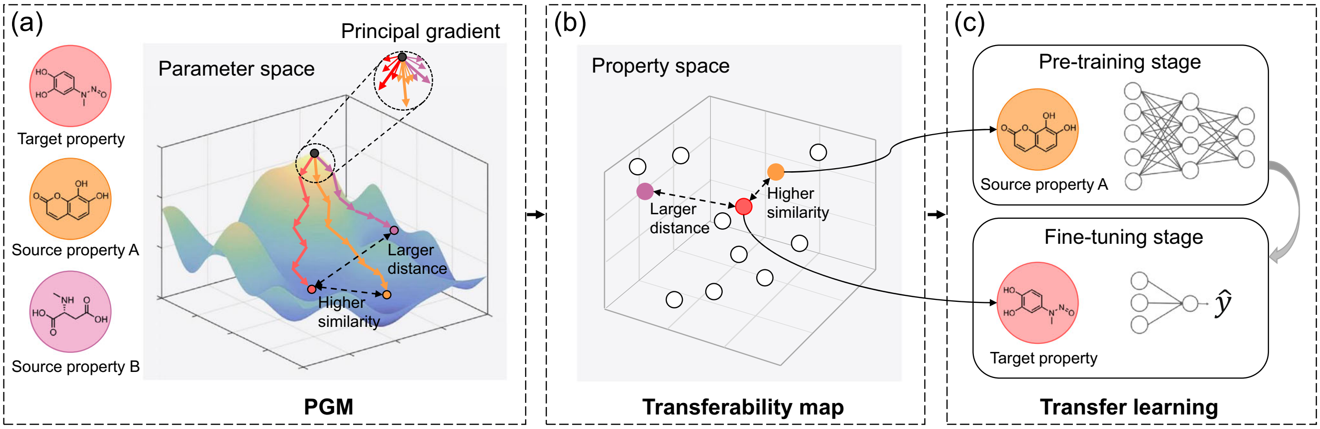


Fig. 1 | Illustration of Principal Gradient-based Measurement (PGM) for guiding transfer learning in molecular property prediction. **a** Firstly, we design an optimization-free scheme to calculate a principal gradient for approximating the direction of model optimization on a molecular property prediction dataset. PGM measures the transferability as the distance between the principal gradient obtained from the source dataset and that derived from the target dataset. **b** Then, we perform PGM on various molecular property prediction datasets to build a quantitative transferability map indicating inter-property correlations. **c** Finally, through the map, we can effectively choose a source dataset with higher similarity to the given target dataset in transfer learning.

calculated as a prior. Thus, we are motivated to pursue a computation-efficient transferability measurement.

Gradients can capture intrinsic task-related characteristics, playing a strong predictive role in model optimization on a dataset, as evidenced by recent studies[43–46]. Grad2Task[43] posits that gradients can be used as features to capture task nature and distinguish between tasks under a meta-learning framework. Numerous gradient surgery techniques[44,45] have been proposed to address gradient conflict in multitask learning. In light of gradients being explored in the context of task-relatedness, we derive PGM, a simple measure based on gradients, to quantify the transferability concerning task-relatedness. Specifically, the proposed principal gradient, obtained through model re-initialization and gradient expectation calculation, approximately indicates the direction for model optimization on a dataset. Afterwards, PGM measures the transferability as the distance between the principal gradient obtained from model training on the source dataset and that derived from the target dataset.

Transferability map building. To visualize the task-relatedness between the source and target molecular property prediction datasets, we create a quantitative transferability map based on the distances between the principal gradients obtained from these datasets, shown in Fig. 1b. When assessing the transferability from potential source datasets to a given target dataset, the distance between source and target datasets serves as an indicator of their knowledge transfer ability. We utilize the PGM distance as the distance metric, which is measured by the difference between the principal gradients of model training on each pair of the datasets.

Application of transfer learning. To validate the effectiveness of the transferability map, we further explore transfer learning experiments in Fig. 1c. Given a target dataset, we select the most similar source dataset for pre-training. During the fine-tuning phase, we fix the feature extractor initialized from the pre-trained model, while train the predictor from scratch on the target task.

The proposed methodology is tested on multiple molecular benchmarks, and the results are described in the following sections.

Transferability map built by PGM

Molecular properties of interest can vary widely in scale, ranging from macroscopic influences on the human body to microscopic electronic properties, such as toxicity to humans[47], the ability of drugs to permeate the brain[48], and hydration free energy[49]. We evaluate the task-relatedness of 12 benchmark datasets from MoleculeNet[42] in three categories: biophysics, physiology, and physical chemistry.

We build a quantitative transferability map to intuitively observe the task-relatedness between these molecular property prediction datasets, shown in Fig. 2. We perform PGM on each dataset to obtain its principal gradient. The pairwise PGM distances between the obtained principal gradients are used to compute the 12 × 12 transferability map, showing the transferability of these datasets in property space distributions.

Transferability map-guided transfer learning improves performance

To investigate the effectiveness of the transferability map in measuring task-relatedness, we design a transferability map-guided cross-task transfer learning strategy. Specifically, each of the 12 datasets is used as the target dataset, while the remaining 11 datasets are employed as source datasets, as described below. Initially, the model is trained on each source dataset to obtain pre-trained models. Subsequently, each of these pre-trained models is fine-tuned on the target dataset, with learning curves shown in Fig. S1 of Supplementary Results. We refer to this experiment as the main experiment below. Due to page limit, the wall-clock time comparison between PGM and fine-tuning is available in Table S2 of Supplementary Results.

As depicted in Fig. 3, a significant correlation between the predicted transferability and the transfer learning performance across various tasks can be observed. As anticipated, when arranging the PGM distance in ascending order, the corresponding transfer performance overall exhibits a gradual decrease for the initial 9 classification tasks (ROC-AUC, higher is better) and a gradual increase for the final 3 regression tasks (RMSE, lower is better). Meanwhile, as the PGM distance increases, there is a decline in transfer performance observed in all 11 scenarios for BACE[50] and more than 8 scenarios for the other classification datasets. For the Tox21 source dataset, the knowledge transfer from ToxCast achieves more improvement in ROC-AUC than any other source dataset. Similarly, for the ToxCast source dataset, Tox21 outperforms 8 out of 10 source datasets in terms of improvement. This is expected because both Tox21 and ToxCast focus on exploring the toxicity of compounds to humans. Another interesting finding is that, among all source datasets, PCBA is a top-three performer in boosting the performance of almost all target datasets in classification scenarios, due to its diverse structure or useful out-of-distribution information. Additionally, despite belonging to distinct domains, some datasets exhibit higher task similarity, such as HIV, MUV, PCBA, Tox21, ToxCast, and SIDER. Collectively, these findings corroborate the utility of the transferability map and suggest that it can effectively assist in selecting the most similar source dataset for the given target dataset, thereby improving the transfer performance.

Transferability map generalizable across subtasks

Considering that the above results verify the effectiveness of the transferability map in measuring similarity across various properties, we next sought to explore whether the transferability map is generalizable across subtasks within these properties. Among the aforementioned three categories of properties, physiology focuses on macroscopic life systems, biophysics uses physical methods to study biological phenomena, while physical chemistry analyzes the principles of the chemical behavior of material systems. Here, we consider knowledge transfer across (1) two different physiology multitask datasets: from Tox21[47] to SIDER[51], and vice versa; (2) one biophysics and one physiology multitask datasets: from MUV[52] to Tox21. Each dataset is associated with several binary classification tasks. We treat each subtask within the target dataset as an individual target dataset and consider the source dataset as a multisource dataset. The pre-training and fine-tuning strategy remains consistent with the main experiment.

The transferability maps between Tox21 and SIDER, as well as between MUV and Tox21, are depicted in Fig. S2 of Supplementary Results. Figure 4 shows the relationship between the PGM distance and the transfer performance in various settings. Specifically, we select three subtasks at regular intervals within each multitask target dataset for analysis. The transfer performance is expected to decrease as the PGM distance increases in all settings. In addition, transferring knowledge within physiology datasets, including Tox21 → SIDER and SIDER → Tox21, consistently shows this trend in over 80% of the target subtasks. Similarly, MUV → Tox21 exhibits the same trend in over 70% of the target subtasks. This observation confirms that knowledge transfer within the same category outperforms transfer between different categories. Overall, it shows that in these cases, the transferability map can be well generalized across the subtasks within the multitask properties.

Ablation study

To gain deeper insights into the effectiveness of each module in PGM, we conduct ablation studies for subsequent analysis: (1) Whether PGM is in a computation-efficient manner? (2) Does the effectiveness of PGM depend on the target dataset size? (3) Does PGM exhibit any relationship with the differences in dataset sizes across diverse tasks? Accordingly, we carry out three experimental groups: (1) Impact of training epochs of PGM, (2) Impact of target dataset size, (3) Fairness comparison on dataset sizes.

Impact of training epochs of PGM. To investigate the impact of the number of reiterative training epochs for computing PGM, we conduct PGM experiments with varying epochs. As shown in Fig. 5, the absolute PGM distance values decrease as the number of training epochs increases

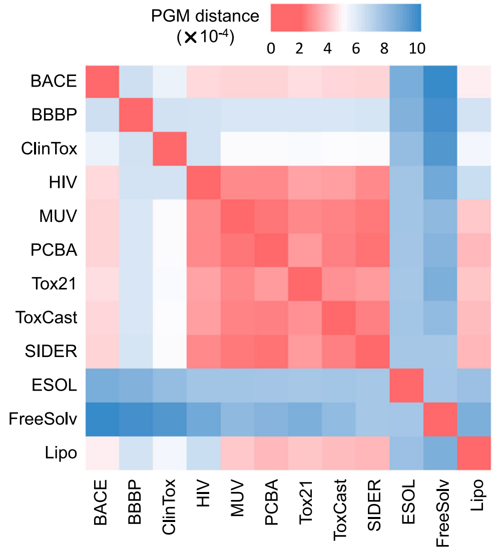


Fig. 2 | Transferability map illustrating the task-relatedness between the 12 benchmark molecular property prediction datasets. Red cells with smaller values indicate higher task similarity, while blue ones with larger values indicate lower task similarity.

from 1 to 10 and then to 20. Despite this decrease, the trends regarding task similarity remain consistent across the three maps. PGM perfectly satisfies our need for computation efficiency, as even one epoch of principle gradient suffices. Based on this observation, we empirically recommend applying PGM on these datasets with a few reiterative training epochs (e.g., around 10) to alleviate the abnormal gradients. Due to page limit, we depicte more representative transferability maps at several key epochs and difference heatmaps in Fig. S3 of Supplementary Results.

Impact of target dataset size. To study the impact of target dataset size, we evaluate the correlation between the PGM distance and the transfer performance with different target dataset sizes. We employ Kendall's τ[53] to assess the correlation between two rankings ordered by transfer performance and PGM distances. The target datasets are generated by randomly sampling subsets from a original target dataset with varying percentages, while the remaining 11 datasets consisting the complete dataset are used as source datasets. Figure 6 shows that the PGM distance generally has a significant linear correlation with the transfer performance, particularly when the target subset size exceeds 40% of the target dataset. The absolute values of Kendall's τ drop less significantly in classification target tasks compared to regression target tasks when the target sample size decreases. This can be attributed to the larger sizes of the classification datasets overall. These findings demonstrate the superior and extensible performance of PGM, serving as a reference standard for transfer learning in molecular property prediction, even with limited target samples.

Fairness comparison on dataset sizes. As for the problem that datasets on different properties vary in dataset sizes, we conduct fairness comparison experiments. Considering the availability of dataset sizes (Table S1 in Supplementary Note 2), we randomly select 8000 molecules from each of the source datasets, namely HIV[54], MUV[52], Tox21 (all 7831 molecules included)[47], and ToxCast[55]. We randomly select 1000 molecules from each of the target datasets, including BACE[50], BBBP[48], ClinTox[56], and SIDER[51].

Similarly, in Fig. 7, the PGM distance generally exhibits a significant linear correlation with the transfer performance, consistent with the main experimental finding. This demonstrates the robustness of PGM, as the correlation between the PGM distance and the transfer performance remains unaffected by variations in data volume across different datasets.

Conclusion

In this study, we propose Principal Gradient-based Measurement (PGM) to support transferability quantification for molecular property prediction datasets. Specifically, we design a principal gradient to approximate model optimization, which performs on source and target datasets to realize transferability mesaure between datasets. Furthermore, we build a transferability map based on PGM to access task-relatedness prior to applying transfer learning. Both theoretical and empirical studies demonstrate that PGM strongly correlates with the transfer performance of molecular property prediction, making it a quantified transferability measure for source dataset selection.

The superiority in predicting the transferability between molecular property prediction datasets reflects the potential of our PGM framework, and there are interesting and promising future works based on PGM. (1) Extending the one-to-one transfer manner to simultaneously selecting the top-n similar source molecular property prediction datasets during the pre-training phase. (2) Improving the transferability metric to guide the design of neural networks and training objectives for molecule generation and optimization tasks.

Methods

Principal Gradient-based Measurement (PGM) for quantifying transferability

Notation and problem definition. Given a target task T_t and a predefined set of source tasks $\{(T_{s(n)}\}_{n=1}^{N}$, the goal of this work is to select the most similar source task for the target task via quantifying their transferability. The corresponding target dataset and source datasets are represented by D_t and $\{(D_{s(n)}\}_{n=1}^{N}$. The model, denoted by $F = g \cdot w$, comprises a feature extractor g and a predictor w. Prior research has demonstrated that higher layers encode more semantic patterns specific to source tasks, while lower-layer features are more generic and related to transferability[57]. Especially when a pair of tasks lack sufficient similarity, transferring higher layers may hurt the performance of a target task. In other words, the transferability is more related to the feature extractor. Thus, we consider the transferability solely focusing on the feature extractor g. We denote the model optimization procedure from the initial point to the optimal convergence point in the parameter space, where the optimal feature extractor g^* is initialized from g_0. Building on the feature extractor, we define the transferability as the distance between the optimal feature extractor for the source task and that for the target task with the metric function $d(\cdot)$:

Definition 1. **Transferability**. The transferability of a feature extractor g from the source task T_s to the target task T_t, denoted by $PGM_{T_s \to T_t}(g)$, is measured by the distance between the optimal feature extractor for the source task $g^*_{T_s}$ and that for the target task $g^*_{T_t}$: $PGM_{T_s \to T_t}(g) = d(g^*_{T_s}, g^*_{T_t})$.

This definition of transferability can be used for selecting a source dataset among $\{D_{s(n)}\}_{n=1}^{N}$ for a given target dataset D_t in transfer learning.

PGM. Based on the feature extractor, the loss function value at the initialized point g_0 and the optimal convergence point g^* are denoted by $\mathcal{L}(g_0)$ and $\mathcal{L}(g^*)$, respectively. We first initialize the model with random weight (g_0, w_0) and make a single forward pass of model training on a dataset to compute the gradient of feature extractor parameters, denoted by $\nabla\mathcal{L}(g_0)$. Second, instead of updating model parameters, we reinitialize the model weights to the same starting point and still perform a single forward pass to collect $\nabla\mathcal{L}(g_0)$. We repeat this process multiple times and then compute the expectation of all the collected gradients, so as to mitigate the potential impact of abnormal gradients and ensure robustness. Finally, we name the expectation of the gradients as the principal gradient, defined by

$$PGM(g_0) = E_{g_0}[\nabla\mathcal{L}(g_0)], \tag{1}$$

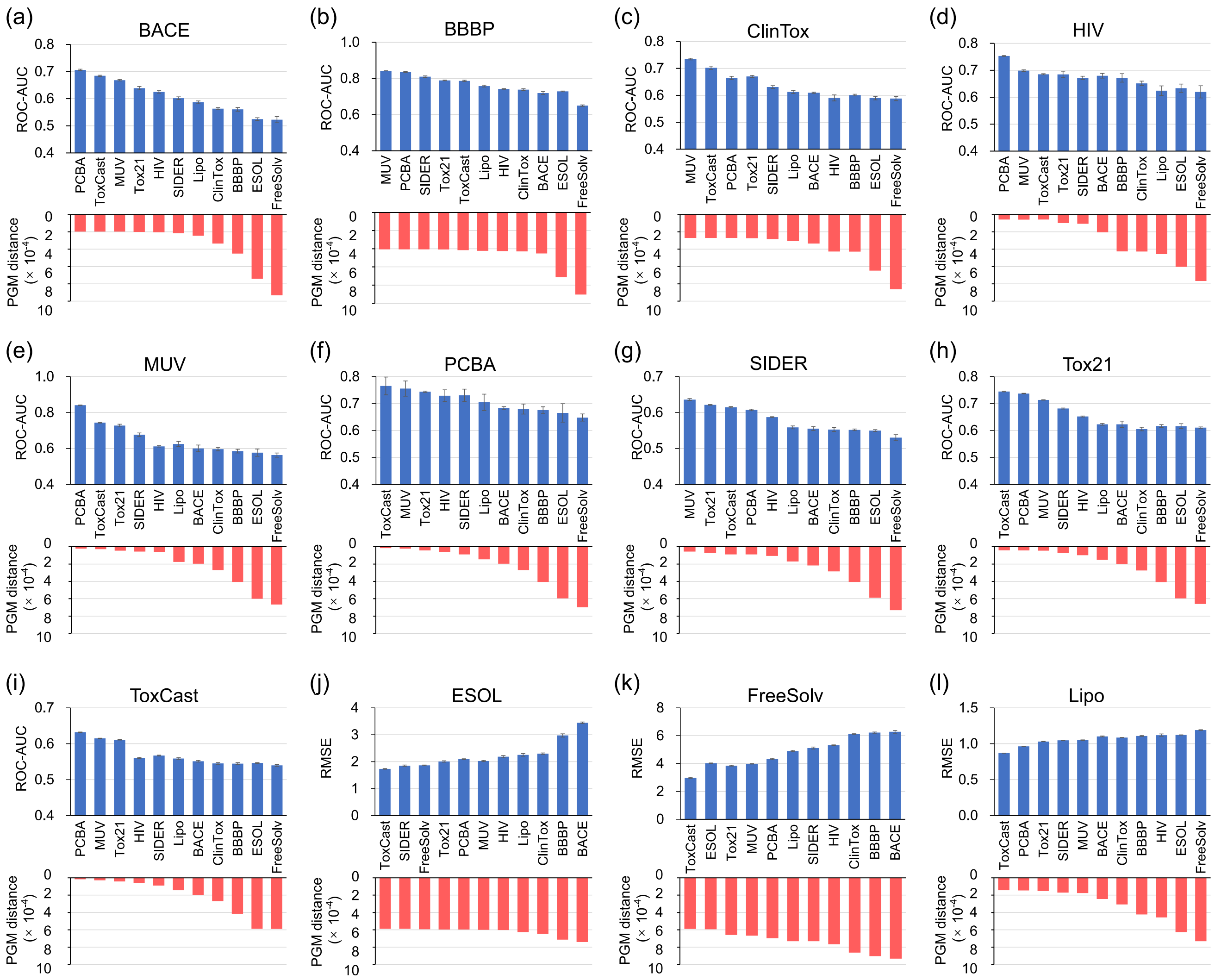


Fig. 3 | Comparison of the PGM distance and the transfer performance on the 12 target datasets. The 12 targets include 9 classification datasets (**a–i**) and 3 regression datasets (**j–l**). For transfer learning experiments, the mean and standard deviation values for five experimental runs are reported.

Transferability quantification by PGM. Model optimization is aptly characterized as a gradient descent procedure, where the model parameters are iteratively adjusted to minimize the loss function. This process can be mathematically formulated utilizing Taylor's formula[58]. To strike a balance between efficiency and effectiveness, we only employ the first-order gradient. Concretely,

$$\mathcal{L}(g^*) = \mathcal{L}(g_0) + \nabla\mathcal{L}(g_0)(g^* - g_0) + \alpha, \tag{2}$$

where α represents the remainder beyond the first-order approximation. From (2), we derive

$$(g^* - g_0) + \frac{\alpha}{\nabla\mathcal{L}(g_0)} = \frac{\mathcal{L}(g^*) - \mathcal{L}(g_0)}{\nabla\mathcal{L}(g_0)}, \tag{3}$$

where $(g^* - g_0)$ denotes the optimization distance between the initial point and the optimal convergence point. However, since the optimal feature extractor g^* is inaccessible without optimization, we need a more simplified and accessible characterization for the distance. As α contains high-order terms, indicating the complexity of the optimization process, we reasonably assume that $(g^* - g_0)$ and $\frac{\alpha}{\nabla\mathcal{L}(g_0)}$ jointly reflect the optimization difficulty of g, which is proportional to $\frac{1}{\nabla\mathcal{L}(g_0)}$. Therefore, we define $\frac{1}{\nabla\mathcal{L}(g_0)}$ as a distance metric to measure the distance from g_0 to g^* as follows:

$$d(g_0, g^*) = \frac{1}{\nabla\mathcal{L}(g_0)}, \tag{4}$$

Next we consider the distance metric on source and target task as

$$\begin{aligned} d_{T_s}(g_0, g^*) &= \tfrac{1}{\nabla\mathcal{L}_{T_s}(g_0)}, \\ d_{T_t}(g_0, g^*) &= \tfrac{1}{\nabla\mathcal{L}_{T_t}(g_0)}. \end{aligned} \tag{5}$$

We define the distance between the optimal feature extractor for the source task $g^*_{T_s}$ and that for the target task $g^*_{T_t}$ by parameter matrix subtraction as

$$\begin{aligned} d(g^*_{T_s}, g^*_{T_t}) &= \| d_{T_s}(g_0, g^*) - d_{T_t}(g_0, g^*)\|_2 \\ &= \| \frac{1}{\nabla\mathcal{L}_{T_s}(g_0)} - \frac{1}{\nabla\mathcal{L}_{T_t}(g_0)} \|_2 \\ &= \frac{\| \nabla\mathcal{L}_{T_t}(g_0) - \nabla\mathcal{L}_{T_s}(g_0)\|_2}{\| \nabla\mathcal{L}_{T_s}(g_0)\, \nabla\mathcal{L}_{T_t}(g_0)\|_2} \\ &\geq \frac{\| \nabla\mathcal{L}_{T_t}(g_0) - \nabla\mathcal{L}_{T_s}(g_0)\|_2}{\| \nabla\mathcal{L}_{T_s}(g_0)\|_2 \;\| \nabla\mathcal{L}_{T_t}(g_0)\|_2}, \end{aligned} \tag{6}$$

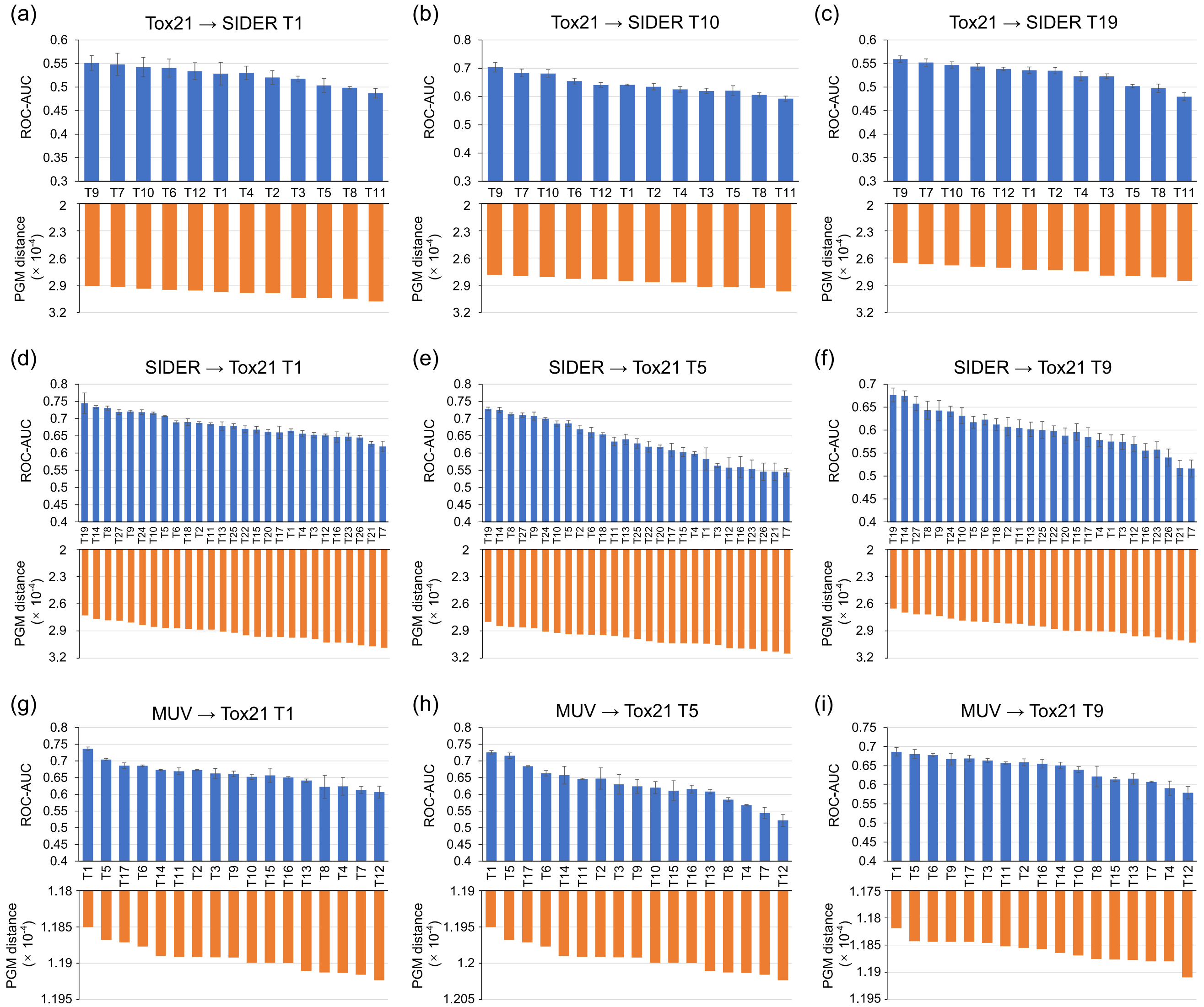


Fig. 4 | Comparison of the PGM distance and the transfer performance across subtasks within different molecular property prediction datasets. The comparison performance on (1) two different physiology multitask datasets: from Tox21 to SIDER (**a**, **b**, **c**), and vice versa (**d**, **e**, **f**); (2) one biophysics and one physiology multitask datasets: from MUV to Tox21 (**g**, **h**, **i**). For each multitask target dataset, we select three subtasks at regular intervals within the dataset as individual target datasets. For transfer learning experiments, the mean and standard deviation values for five experimental runs are reported.

where the inequality is derived from Cauchy-Schwarz inequality[59].

As calculating the gradient only once may lead to abnormal value, we apply PGM by computing the expectation of gradients in (6):

$$PGM_{T_s \to T_t}(g) = d(g^*_{T_s}, g^*_{T_t}) = \frac{\|E_{g_0}[\nabla \mathcal{L}_{T_t}(g_0)] - E_{g_0}[\nabla \mathcal{L}_{T_s}(g_0)]\|_2}{\|E_{g_0}[\nabla \mathcal{L}_{T_s}(g_0)]\|_2 \, \|E_{g_0}[\nabla \mathcal{L}_{T_t}(g_0)]\|_2}, \tag{7}$$

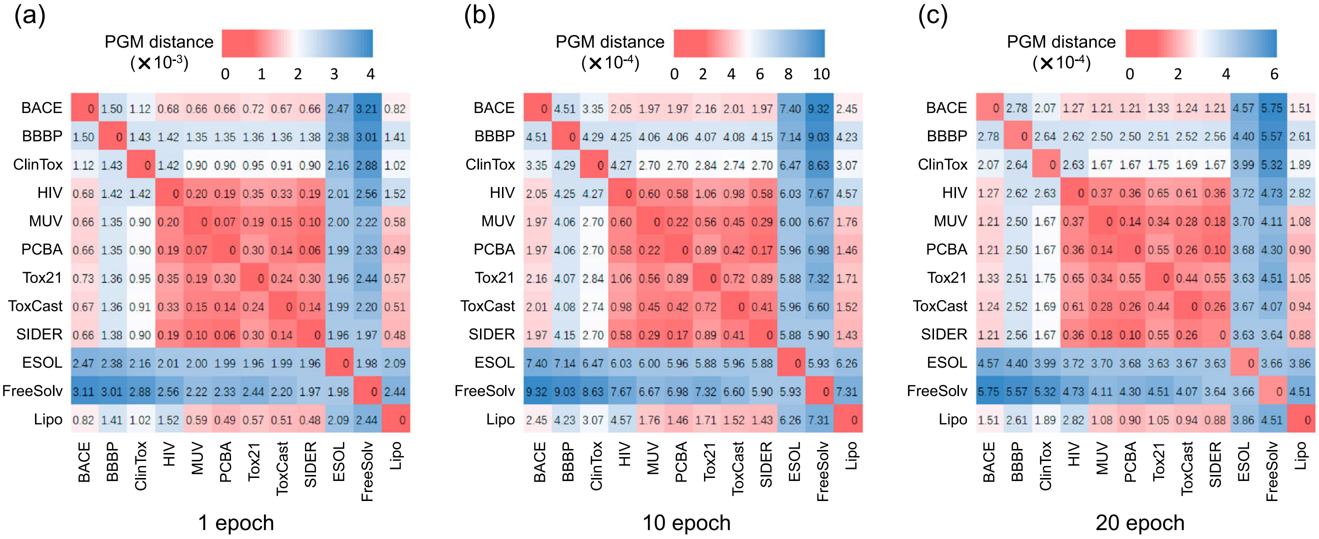
The distance metric measures the gap between $g^*_{T_s}$ and $g^*_{T_t}$. Finally, we use $d(g^*_{T_s}, g^*_{T_t})$ named PGM distance to denote the transferability from the source task to the target task. The above process of quantifying transferability with PGM is efficient, as the only computation is repeatedly making a single forward pass of model training on datasets without optimization. Further algorithm details are in Supplementary Note 1.

Transferability map-guided transfer learning

Based on PGM, we develop a quantitative transferability map to provide a panoramic view of task-relatedness between the source and target molecular property prediction datasets. Specifically, we employ the pairwise PGM distances as the transferability metric between these datasets. In the main experiment conducted on 12 benchmark datasets from MoleculeNet[42], we utilize the 12 datasets to build the transferability map. In the experiment generalizable across subtasks, we use all the subtasks within corresponding molecular property prediction datasets to build the transferability maps. With the guidance of the transferability maps, we perform transfer learning in various settings and for different tasks, which can be found in Results. Finally, We evaluate the correlation between the predicted transferability by PGM and the transfer performance.

Experimental setup

Datasets. We use 12 benchmark datasets from MoleculeNet[42], including 9 binary classification tasks and 3 regression tasks. The datasets cover molecular data from a wide range of domains, such as physical chemistry, biophysics, and physiology. In PGM, We conduct experiments on all datasets using the complete dataset. In pre-training and fine-tuning, we perform five independent runs on five random-seeded scaffold splitting for all datasets with a train/validation/test ratio of 8:1:1, as suggested by the MoleculeNet[42]. Scaffold splitting[60] splits molecules based on their scaffolds (molecular substructures), which can better evaluate the generalization ability of the models on out-of-distribution data samples.



(a) 1 epoch

	BACE	BBBP	ClinTox	HIV	MUV	PCBA	Tox21	ToxCast	SIDER	ESOL	FreeSolv	Lipo
BACE	0	1.50	1.12	0.68	0.66	0.66	0.72	0.67	0.66	2.47	3.21	0.82
BBBP	1.50	0	1.43	1.42	1.35	1.35	1.36	1.36	1.38	2.38	3.01	1.41
ClinTox	1.12	1.43	0	1.42	0.90	0.90	0.95	0.91	0.90	2.16	2.88	1.02
HIV	0.68	1.42	1.42	0	0.20	0.19	0.35	0.33	0.19	2.01	2.56	1.52
MUV	0.66	1.35	0.90	0.20	0	0.07	0.19	0.15	0.10	2.00	2.22	0.58
PCBA	0.66	1.35	0.90	0.19	0.07	0	0.30	0.14	0.06	1.99	2.33	0.49
Tox21	0.73	1.36	0.95	0.35	0.19	0.30	0	0.24	0.30	1.96	2.44	0.57
ToxCast	0.67	1.36	0.91	0.33	0.15	0.14	0.24	0	0.14	1.99	2.20	0.51
SIDER	0.66	1.38	0.90	0.19	0.10	0.06	0.30	0.14	0	1.96	1.97	0.48
ESOL	2.47	2.38	2.16	2.01	2.00	1.99	1.96	1.99	1.96	0	1.98	2.09
FreeSolv	3.11	3.01	2.88	2.56	2.22	2.33	2.44	2.20	1.97	1.98	0	2.44
Lipo	0.82	1.41	1.02	1.52	0.59	0.49	0.57	0.51	0.48	2.09	2.44	0

(b) 10 epoch

	BACE	BBBP	ClinTox	HIV	MUV	PCBA	Tox21	ToxCast	SIDER	ESOL	FreeSolv	Lipo
BACE	0	4.51	3.35	2.05	1.97	1.97	2.16	2.01	1.97	7.40	9.32	2.45
BBBP	4.51	0	4.29	4.25	4.06	4.06	4.07	4.08	4.15	7.14	9.03	4.23
ClinTox	3.35	4.29	0	4.27	2.70	2.70	2.84	2.74	2.70	6.47	8.63	3.07
HIV	2.05	4.25	4.27	0	0.60	0.58	1.06	0.98	0.58	6.03	7.67	4.57
MUV	1.97	4.06	2.70	0.60	0	0.22	0.56	0.45	0.29	6.00	6.67	1.76
PCBA	1.97	4.06	2.70	0.58	0.22	0	0.89	0.42	0.17	5.96	6.98	1.46
Tox21	2.16	4.07	2.84	1.06	0.56	0.89	0	0.72	0.89	5.88	7.32	1.71
ToxCast	2.01	4.08	2.74	0.98	0.45	0.42	0.72	0	0.41	5.96	6.60	1.52
SIDER	1.97	4.15	2.70	0.58	0.29	0.17	0.89	0.41	0	5.88	5.90	1.43
ESOL	7.40	7.14	6.47	6.03	6.00	5.96	5.88	5.96	5.88	0	5.93	6.26
FreeSolv	9.32	9.03	8.63	7.67	6.67	6.98	7.32	6.60	5.90	5.93	0	7.31
Lipo	2.45	4.23	3.07	4.57	1.76	1.46	1.71	1.52	1.43	6.26	7.31	0

(c) 20 epoch

	BACE	BBBP	ClinTox	HIV	MUV	PCBA	Tox21	ToxCast	SIDER	ESOL	FreeSolv	Lipo
BACE	0	2.78	2.07	1.27	1.21	1.21	1.33	1.24	1.21	4.57	5.75	1.51
BBBP	2.78	0	2.64	2.62	2.50	2.50	2.51	2.52	2.56	4.40	5.57	2.61
ClinTox	2.07	2.64	0	2.63	1.67	1.67	1.75	1.69	1.67	3.99	5.32	1.89
HIV	1.27	2.62	2.63	0	0.37	0.36	0.65	0.61	0.36	3.72	4.73	2.82
MUV	1.21	2.50	1.67	0.37	0	0.14	0.34	0.28	0.18	3.70	4.11	1.08
PCBA	1.21	2.50	1.67	0.36	0.14	0	0.55	0.26	0.10	3.68	4.30	0.90
Tox21	1.33	2.51	1.75	0.65	0.34	0.55	0	0.44	0.55	3.63	4.51	1.05
ToxCast	1.24	2.52	1.69	0.61	0.28	0.26	0.44	0	0.26	3.67	4.07	0.94
SIDER	1.21	2.56	1.67	0.36	0.18	0.10	0.55	0.26	0	3.63	3.64	0.88
ESOL	4.57	4.40	3.99	3.72	3.70	3.68	3.63	3.67	3.63	0	3.66	3.86
FreeSolv	5.75	5.57	5.32	4.73	4.11	4.30	4.51	4.07	3.64	3.66	0	4.51
Lipo	1.51	2.61	1.89	2.82	1.08	0.90	1.05	0.94	0.88	3.86	4.51	0

Fig. 5 | Impact of training epochs of PGM. Transferability maps of performing PGM with (**a**) 1, (**b**) 10, and (**c**) 20 epochs to explore the minimum number of epochs required for training. Red cells with smaller values indicate higher task similarity, while blue ones with larger values indicate lower task similarity.

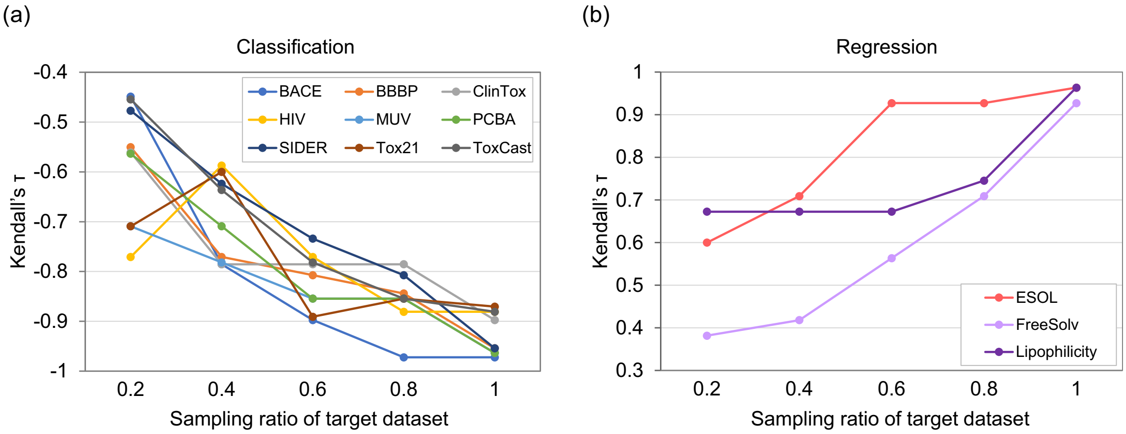


Fig. 6 | Impact of target dataset size. Impact of the sampling ratio of target datasets to the performance of PGM, when transferring the remaining 11 source datasets to each target dataset in both **a** classification and **b** regression tasks.

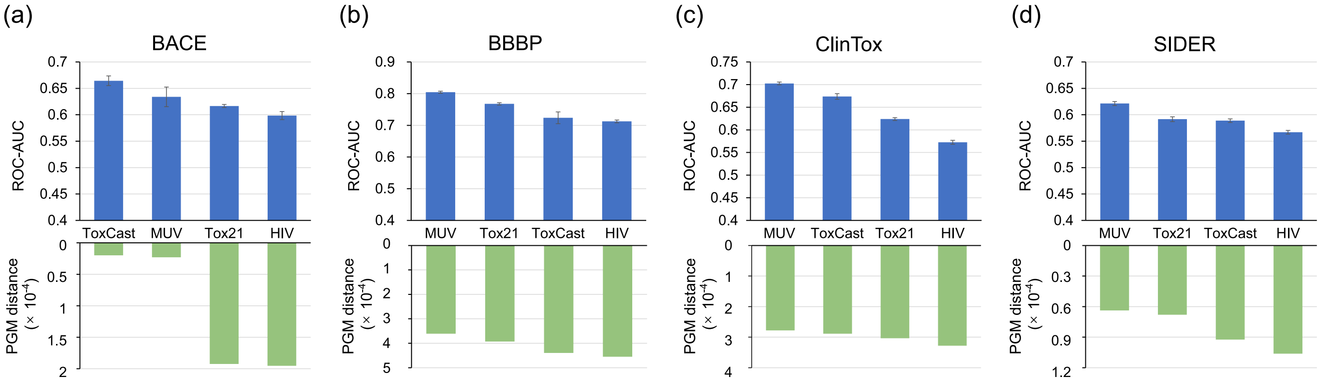


Fig. 7 | Impact of source and target dataset size to the performance of PGM. Impact evaluated by comparing the PGM distance and the transfer performance on transferring the four source datasets (HIV, MUV, Tox21, and ToxCast), each containing 8000 molecules, to each of the target datasets (**a** BACE, **b** BBBP, **c** ClinTox, and **d** SIDER), each consisting of 1000 molecules. For transfer learning experiments, the mean and standard deviation values for five experimental runs are reported.

Table S1 in Supplementary Note 2 contains more details about the datasets.

Implementation details. We use the open-source chemical analysis tool RDKit[61] to convert raw SMILES into 2D molecular graphs and extract atom features (atom number and chirality tag) and bond features (bond type and bond direction). We adopt graph isomorphism network (GIN)[62] based on DGL-LifeSci package[63] to extract the molecular graph representation in both PGM and transfer learning experiments. GIN is sensitive to the number of layers, and we employ a widely used 3-layer setting. We also use the average pooling as the READOUT function to obtain the graph representation. We adopt a single-layer MLP as the property prediction network.

In PGM, we empirically perform model training for 10 epochs on various datasets. To build the transferability map, we compute pairwise PGM distances between the principal gradient matrices of these datasets. Specifically, we extract the weight parameters from these gradient matrices and concatenate them into a single tensor. Subsequently, we calculate the pairwise PGM distances between these concatenated tensors.

In the transfer learning experiments, during pre-training, we use the Adam optimizer with a learning rate of 1×10^{-3}, optimizing BCEWithLogitsLoss for classification and SmoothL1Loss for regression tasks. We train pre-trained models on each dataset with a batch size of 32 and 200 epochs, and select the one with the best performance on the validation metric. This ensures our pre-trained models are well-trained and generalize effectively to new data. In fine-tuning, we initialize the feature extractor from the pre-trained model and train the predictor from scratch on the target task. We fine-tune five times with a batch size of 32 to report the average and standard deviation of performance on the testing set, using ROC-AUC for classification and RMSE for regression tasks.

PGM is implemented utilizing Pytorch and runs on an Ubuntu Server with NVIDIA GeForce RTX 3090Ti graphics processing units.

Evaluation metrics. As suggested by the MoleculeNet[42], we use ROC-AUC as the evaluation metric for the binary classification datasets, for which higher is better. With respect to regression datasets, we use RMSE, for which lower is better. We perform five independent runs with five random seeds for each method and report the mean and the standard deviation of the metrics.

In the ablation study, we measure the similarity between two rankings ordered by transfer performance and PGM distances using Kendall's τ[53], also known as Kendall rank correlation coefficient. Kendall's τ value ranges from -1 to 1, with higher values indicating consistent rankings, lower values implying reversed rankings, and 0 suggesting unrelated rankings.

Data availability

The datasets used in our paper are publicly available on the MoleculeNet website https://moleculenet.org/datasets-1.

Code availability

Codes for training the source data and analyzing the results are available on the Zenodo at https://doi.org/10.5281/zenodo.10071500.

Received: 5 November 2023; Accepted: 5 April 2024;
Published online: 17 April 2024

References

1. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
2. Shen, J. & Nicolaou, C. A. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discov. Today.: Technol.* **32**, 29–36 (2019).
3. Nigam, A. et al. Tartarus: A benchmarking platform for realistic and practical inverse molecular design. *Advances in Neural Information Processing Systems***36** (2024).
4. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32**, 40–51 (2014).
5. Dowden, H. & Munro, J. Trends in clinical success rates and therapeutic focus. *Nat. Rev. Drug Discov.* **18**, 495–496 (2019).
6. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *J. Big Data* **3**, 1–40 (2016).
7. Kenton, J. D. M.-W. C. & Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 4171–4186 (2019).
8. Brown, T. et al. Language models are few-shot learners. *Adv. neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
9. Gidaris, S., Singh, P. & Komodakis, N. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations* (2018).
10. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660 (2021).
11. Fang, X. et al. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* **4**, 127–134 (2022).
12. Heydari, S., Raniolo, S., Livi, L. & Limongelli, V. Transferring chemical and energetic knowledge between molecular systems with machine learning. *Commun. Chem.* **6**, 13 (2023).
13. Flam-Shepherd, D., Zhu, K. & Aspuru-Guzik, A. Language models can learn complex molecular distributions. *Nat. Commun.* **13**, 3293 (2022).
14. Skinnider, M. A., Stacey, R. G., Wishart, D. S. & Foster, L. J. Chemical language models enable navigation in sparsely populated chemical space. *Nat. Mach. Intell.* **3**, 759–770 (2021).
15. Zaverkin, V., Holzmüller, D., Bonfirraro, L. & Kästner, J. Transfer learning for chemically accurate interatomic neural network potentials. *Phys. Chem. Chem. Phys.* **25**, 5383–5396 (2023).
16. Chithrananda, S., Grand, G. & Ramsundar, B. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *Machine Learning for Molecules Workshop, NeurIPS* (2020). Preprint at https://arxiv.org/abs/2010.09885.
17. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
18. Bai, R. et al. Transfer learning: making retrosynthetic predictions based on a small chemical reaction dataset scale to a new level. *Molecules* **25**, 2357 (2020).
19. Pesciullesi, G., Schwaller, P., Laino, T. & Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nat. Commun.* **11**, 4874 (2020).
20. Grambow, C. A., Pattanaik, L. & Green, W. H. Deep learning of activation energies. *J. Phys. Chem. Lett.* **11**, 2992–2997 (2020).
21. Espley, S. G., Farrar, E. H., Buttar, D., Tomasi, S. & Grayson, M. N. Machine learning reaction barriers in low data regimes: a horizontal and diagonal transfer learning approach. *Digit. Discov.* **2**, 941–951 (2023).
22. Singh, S. & Sunoj, R. B. A transfer learning protocol for chemical catalysis using a recurrent neural network adapted from natural language processing. *Digit. Discov.* **1**, 303–312 (2022).
23. Hu, W. et al. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)* (2020).
24. Cai, C. et al. Transfer learning for drug discovery. *J. Med. Chem.* **63**, 8683–8694 (2020).
25. Li, X. & Fourches, D. Inductive transfer learning for molecular activity prediction: Next-gen qsar models with molpmofit. *J. Cheminf.* **12**, 1–15 (2020).

26. Wang, Z., Dai, Z., Póczos, B. & Carbonell, J. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11293–11302 (2019).
27. Zhang, W., Deng, L., Zhang, L. & Wu, D. A survey on negative transfer. *IEEE/CAA J. Autom. Sin.* **10**, 305–329 (2022).
28. Nigam, A. et al. Assigning confidence to molecular property prediction. *Expert Opin. Drug Discov.* **16**, 1009–1023 (2021).
29. Bao, Y. et al. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing (ICIP)*, 2309–2313 (IEEE, 2019).
30. Nguyen, C., Hassner, T., Seeger, M. & Archambeau, C. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, 7294–7305 (PMLR, 2020).
31. You, K., Liu, Y., Wang, J. & Long, M. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, 12133–12143 (PMLR, 2021).
32. Huang, L.-K., Huang, J., Rong, Y., Yang, Q. & Wei, Y. Frustratingly easy transferability estimation. In *International Conference on Machine Learning*, 9201–9225 (PMLR, 2022).
33. Bajusz, D., Rácz, A. & Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **7**, 1–13 (2015).
34. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminf.* **8**, 1–14 (2016).
35. Girschick, T., Rückert, U. & Kramer, S. Adapted transfer of distance measures for quantitative structure-activity relationships and data-driven selection of source datasets. *Comput. J.* **56**, 274–288 (2013).
36. Zamir, A. R. et al. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3712–3722 (2018).
37. Dwivedi, K. & Roig, G. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12387–12396 (2019).
38. Li, H. et al. Improving molecular property prediction through a task similarity enhanced transfer learning strategy. *Iscience* **25**, 105231 (2022).
39. Yao, S. et al. Chemical property relation guided few-shot molecular property prediction. In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE, 2022).
40. Hu, F. et al. A multimodal protein representation framework for quantifying transferability across biochemical downstream tasks. *Adv. Sci.* **10**, 2301223 (2023).
41. Lim, S. et al. A review on compound-protein interaction prediction methods: data, format, representation and model. *Comput. Struct. Biotechnol. J.* **19**, 1541–1556 (2021).
42. Wu, Z. et al. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
43. Wang, J., Wang, K.-C., Rudzicz, F. & Brudno, M. Grad2task: Improved few-shot text classification using gradients for task representation. *Adv. Neural Inf. Process. Syst.* **34**, 6542–6554 (2021).
44. Chen, Z., Badrinarayanan, V., Lee, C.-Y. & Rabinovich, A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, 794–803 (PMLR, 2018).
45. Yu, T. et al. Gradient surgery for multi-task learning. *Adv. Neural Inf. Process. Syst.* **33**, 5824–5836 (2020).
46. Pan, Z. et al. Fedmdfg: Federated learning with multi-gradient descent and fair guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 9364–9371 (2023).
47. Hartung, T. Toxicology for the twenty-first century. *Nature* **460**, 208–212 (2009).
48. Martins, I. F., Teixeira, A. L., Pinheiro, L. & Falcao, A. O. A bayesian approach to in silico blood-brain barrier penetration modeling. *J. Chem. Inf. Model.* **52**, 1686–1697 (2012).
49. Mobley, D. L. & Guthrie, J. P. Freesolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput.-aided Mol. Des.* **28**, 711–720 (2014).
50. Subramanian, G., Ramsundar, B., Pande, V. & Denny, R. A. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *J. Chem. Inf. Model.* **56**, 1936–1949 (2016).
51. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The sider database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075–D1079 (2016).
52. Rohrer, S. G. & Baumann, K. Maximum unbiased validation (muv) data sets for virtual screening based on pubchem bioactivity data. *J. Chem. Inf. Model.* **49**, 169–184 (2009).
53. Kendall, M. G. A new measure of rank correlation. *Biometrika* **30**, 81–93 (1938).
54. Riesen, K. & Bunke, H. Iam graph database repository for graph based Iam graph database repository for graph based. In *SSPR/SPR* Vol. 5342, 287–297 (Springer, 2008).
55. Richard, A. M. et al. Toxcast chemical landscape: paving the road to 21st century toxicology. *Chem. Res. Toxicol.* **29**, 1225–1251 (2016).
56. Gayvert, K. M., Madhukar, N. S. & Elemento, O. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem. Biol.* **23**, 1294–1301 (2016).
57. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? *Advances in neural information processing systems***27** (2014).
58. Kline, M. *Mathematical Thought from Ancient to Modern Times: Volume 2*, vol. 2 (Oxford university press, 1990).
59. Steele, J. M.*The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities* (Cambridge University Press, 2004).
60. Ramsundar, B., Eastman, P., Walters, P. & Pande, V. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more* (" O'Reilly Media, Inc.", 2019).
61. Landrum, G. et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg. Landrum* **8**, 31 (2013).
62. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations* (2019).
63. Li, M. et al. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *ACS Omega* **6**, 27233–27238 (2021).

Acknowledgements

This work is supported by the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (Grant No. SN-ZJU-SIAS-001), Zhejiang Province High-Level Talents Special Support Program "Leading Talent of Technological Innovation of Ten-Thousands Talents Program" (No. 2022R52046), Zhejiang Provincial Science and Technology Project for Public Welfare (LGG22F020007), and Scientific Research Fund of Zhejiang University (No. XY2023020).

Author contributions

S.Y. proposed the research, conducted experiments, analyzed the data, and wrote the manuscript. L.J. and Z.Z. contributed technical support in conducting experiments. J.S., L.C. and M.S. provided evaluation and suggestions. Z.F. supervised the overall project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42004-024-01169-4.

Correspondence and requests for materials should be addressed to Zunlei Feng.

Peer review information *Communications Chemistry* thanks Elliot Farrar and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024