Multi-task Extension of Geometrically Aligned Transfer Encoder

Sung Moon Ko [1 *] **Sumin Lee** [1 *] **Dae-Woong Jeong** [1 *] **Hyunseung Kim** [1] **Chanhui Lee** [1] **Soorin Yim** [1] **Sehui Han** [1]

Abstract

Molecular datasets often suffer from a lack of data. It is well-known that gathering data is difficult due to the complexity of experimentation or simulation involved. Here, we leverage mutual information across different tasks in molecular data to address this issue. We extend an algorithm that utilizes the geometric characteristics of the encoding space, known as the Geometrically Aligned Transfer Encoder (GATE), to a multi-task setup. Thus, we connect multiple molecular tasks by aligning the curved coordinates onto locally flat coordinates, ensuring the flow of information from source tasks to target data to support the performance.

1. Introduction

The quantity of data is a crucial factor in machine learning. However, it is not always feasible to acquire the necessary amount of data in practice. Many efforts have been made to address the data issue. One direct approach is data generation, which aims to generate plausible data (such as through reference augmentations or generation). Another approach is transfer learning, which is more indirect as it leverages mutual information from different source tasks (Zhuang et al., 2011; Long et al.; Zhuang et al., 2013; 2014; Pan et al., 2020; Quattoni et al., 2008; Kulis et al., 2011; Raghu et al., 2019; Yu et al., 2022; Wang et al., 2019; Peng et al., 2021). Lastly, there is multi-task learning, which shares a latent space across given tasks (Caruana, 1997; Zhang & Yang, 2018; Liu et al., 2022; Allenspach et al., 2024).

Despite these achievements, the data issue remains particularly pronounced in scientific endeavors. Scientific experiments or simulations often require significant amounts of time and effort, making it challenging to amass abundant data in the field. Since, our main focus is on molecular property prediction tasks (Scarselli et al., 2009; Bruna et al., 2013; Duvenaud et al., 2015; Defferrard et al., 2016; Jin et al., 2018; Coley et al., 2019; Ko et al., 2023a), we aim to address this issue by utilizing various molecular property datasets.

Our starting point is a transfer algorithm, namely the Geometrically Aligned Transfer Encoder (GATE), which is based on differential geometry (Ko et al., 2023b). This algorithm utilizes the concept of curved geometry in a Riemannian scheme. The key idea of this algorithm is to align the geometrical shapes of the underlying latent spaces of source and target tasks. In general, it is extremely complicated to compute their geometrical characteristics analytically. However, the algorithm bypasses this issue by introducing one crucial mathematical characteristic of Riemannian geometry: diffeomorphism invariance, which guarantees the freedom of coordinate choices at any point on the manifold. Additionally, it ensures that one can always find a locally flat frame under any circumstances. If one can find a locally flat frame over any task, then it is possible to impose a constraint that restricts the geometric shape of coordinates over source and target tasks. If the underlying geometry can be matched, the mutual information across tasks will flow to one another and support model performance on the target task side. However, GATE is proven to work in a two-task setting, with one target and one source task. Yet, theoretically, it is not restricted to two tasks. Therefore, we extend the concept of GATE to multiple sources.

The fundamental concept remains unchanged. Since most molecular properties can be effectively computed from a common representation called SMILES (Weininger, 1988), it is natural to assume that there exists a common manifold for any tasks in molecular property prediction. Since this manifold is curved, imposing constraints to match the shapes of geometries of tasks requires a mapping from task coordinates to their corresponding locally flat frames. With multiple source tasks now present, it is mandatory to find mapping functions over task spaces for each one, as shown in Figure 1. This amplifies the leveraging effect of GATE, as mutual information now flows not only from one source task but also from multiple other sources.

We established an experimental setup based on the extended

[*]Equal contribution [1]LG AI Research, Seoul, Republic of Korea. Correspondence to: Sehui Han <hansse.han@lgresearch.ai>.

Accepted at AI4Science Workshop at the 41 st International Conference on Machine Learning, Vienna, Austria, Vienna, Austria. PMLR 235, 2024.

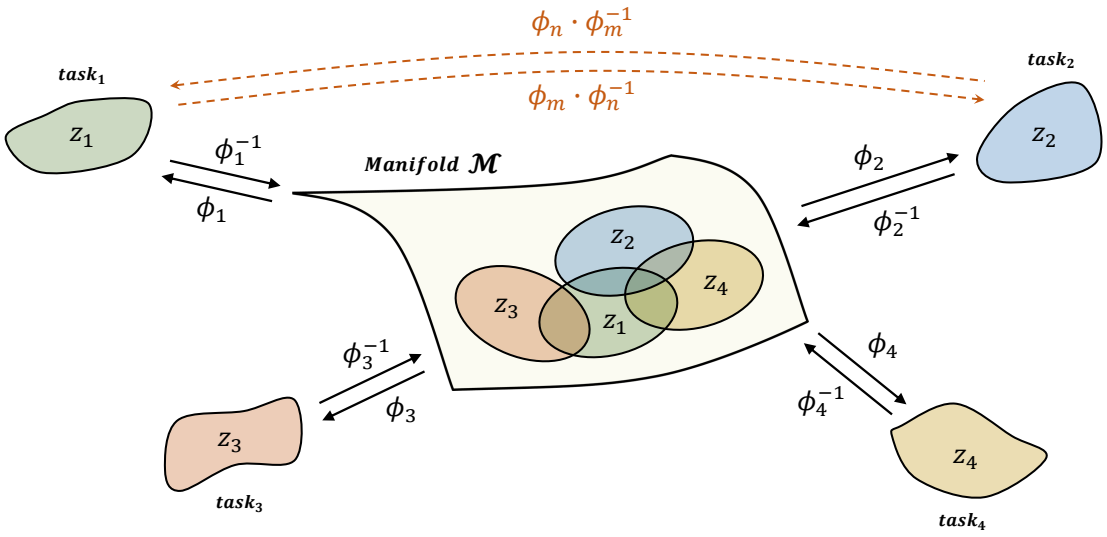

Figure 1. Four different coordinate frames are demonstrated in the figure, with coordinate transformation maps to each pair of tasks. One can interpret each coordinate frame as task-specific coordinates and map them with transfer models. An arbitrary point in the overlapping region of the manifold can be transformed from one task coordinate to another by combining mapping functions ϕ. Moreover, by introducing perturbation points, one can define the distance between points to match the geometrical shape in the overlapping region.

GATE algorithm with multiple molecular property prediction regression tasks from a number of different sources. We have shown that the extended GATE outperforms conventional multi-task learning schemes in terms of performance. Additionally, we conducted ablation test to demonstrate that our algorithm is robust and reliable in multiple combinations of tasks.

Our main contribution of the article is as follows.

- We extend the *GATE* to encode multiple source tasks setup.
- Extension to multiple tasks provides a positive leveraging effect.
- Proposed model outperforms conventional method in multi-task molecular property setup.

2. Multi-Task Extension of Geometrically Aligned Transfer Encoder

Since the latent vector is believed to capture the essence of information for a given task, it is crucial to understand the geometrical characteristics of the latent spaces where the latent vector resides. If two different tasks share common factors in their property inference processes, then one may assume that the geometrical shapes of their latent spaces should be similar. Therefore, if one can align the geometrical shapes of tasks, mutual information will flow through mapping functions, thereby supporting the performance of the target task.

Here, we utilize the GATE algorithm and aim to extend its architecture to accommodate multiple source tasks.[1]

In Figure 2 we first take an input SMILES and embed it into the corresponding vector. After embedding, latent space is formulated by encoders, which consist of DMPNN(Yang et al., 2019) and MLP layers. The latent vector is fed into task-corresponding heads for inference properties. Here we utilize MSE for basic regression loss in the training scheme as follows:

$$l_{\text{reg}} = \frac{1}{N}\sum_{i}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

Where N, y_i, and $\hat{y}_i$ represent the number of data points, target, and predicted value, respectively. The difference now is that there exist multiple tasks, hence, there are also multiple instances of the regression loss.

To align the geometrical shapes of tasks, it is necessary to establish a mapping relation between the latent space and the locally flat frame of the universal manifold. The coordinate mapping can be induced by a Jacobian at an arbitrary point:

$$z'^i \equiv \sum_j \frac{\partial z'^i}{\partial z^j} z^j \tag{2}$$

[1]For basic assumptions and detailed explanation of GATE, refer to (Ko et al., 2023b).

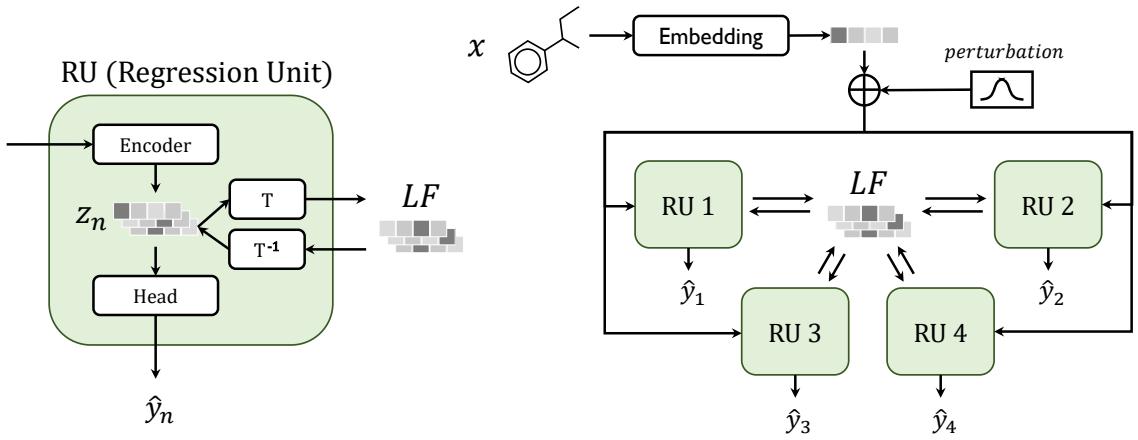

Figure 2. Schematic diagram for the Extended *GATE* algorithm. The algorithm consists of a number of Regression Units. Each Regression Unit corresponds to an individual task. The universal manifold covers the entire coordinate space of RU's, and the transfer model T maps a vector from each RU to a locally flat frame on the universal manifold. One can take the reverse path from the manifold to reconstruct the original vector. Furthermore, one can also transfer a vector to another RU coordinate by utilizing a different task's inverse transfer module.

The model should always be able to differentiate in order to learn via a gradient descent scheme. Hence, we design a mapping function with an autoencoder model. The encoder indicates mapping from latent space to universal manifold, and the decoder indicates mapping the other way around.

$$z'_\alpha = \text{Transfer}_{\alpha\to LF}(z_\alpha) \tag{3}$$
$$\hat{z}_\alpha = \text{Transfer}^{-1}{}_{LF\to\alpha}(z'_\alpha) \tag{4}$$
$$z'_t = \text{Transfer}_{t\to LF}(z_t) \tag{5}$$
$$\hat{z}_t = \text{Transfer}^{-1}{}_{LF\to t}(z'_t) \tag{6}$$

Where t and α indicate the target task and source number of tasks, respectively. If there are k numbers of source tasks, the Greek alphabet runs from $1 \sim k$, and numbers indicate the source task number. For instance, $\text{Transfer}_{t\to LF}(z_t)$ means transformation from target latent to universal manifold and $\text{Transfer}_{5\to LF}(z_5)$ means transformation from source task number 5 to universal manifold. We indeed utilize MSE loss for the autoencoder which consists of transfer and its inverse modules.

$$l_{\text{auto}} = \sum_\alpha \text{MSE}(z_\alpha, \hat{z}_\alpha) \tag{7}$$

Now, everything is set to match the geometrical shapes of latent spaces. Since the encoder maps the latent vector on the latent space to a locally flat frame on the universal manifold, it is straightforward to impose a constraint that matches the latent vector from the target task and the source task. To define the consistency loss, we should recall the definition of the transfer model from the equations mentioned in 3 and 5. As depicted in the equations, $\text{Model}_{t\to LF}$ and $\text{Model}_{\alpha\to LF}$ indicate a model from the target to the locally flat (LF) frame and from the source to the LF frame, respectively. Here, we can impose a series of constraints to align the geometrical shapes from the source and target. One of these constraints requires that the latent vectors from the source and target should have the same value on the universal manifold. This constraint is referred to as the consistency loss.

$$l_{cons} = \sum_\alpha \text{MSE}(z'_\alpha, z'_t) \tag{8}$$

This loss equalizes the target latent and source latent vectors in a locally flat frame on the universal manifold. The latent spaces are also aligned by latent vectors. Furthermore, one can induce another form of constraint to maximize the alignment of latent spaces.

$$z'_\alpha = \text{Transfer}_{\alpha\to LF}(z_\alpha) \tag{9}$$
$$\hat{z}_{\alpha\to t} = \text{Transfer}^{-1}{}_{LF\to t}(z'_\alpha) \tag{10}$$

The equation above illustrates the transformation of a latent vector from the source task to the target task. If the universal manifold is well-defined and both latent spaces from the source and target tasks are aligned properly, then a latent vector transformed from the source to the target task and a latent vector from the target task induced by the same SMILES input should always be the same. Hence, it is straightforward to imagine the specific form of the constraint which is written as follows.

$$l_{map} = \sum_\alpha \text{MSE}(y_t, \hat{y}_{\alpha\to t}) \tag{11}$$

Here, y_t represents the label for the target predicted value, and $\hat{y}_{\alpha\to t}$ indicates the predicted value from $\hat{z}_{\alpha\to t}$. The above loss ensures mutual information flow by aligning locally flat coordinates on the given latent vectors.

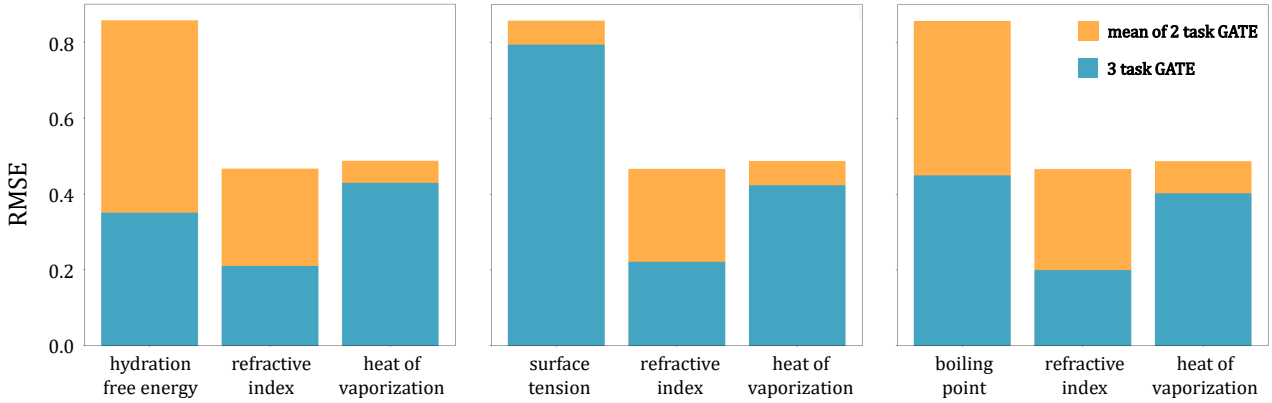

Figure 3. Regression performance of three-task *GATE* and two-task *GATE* in root mean square error (RMSE). For evaluating the regression performance of two-task *GATE*, all three possible pairs of three tasks were trained separately and averaged.

Unfortunately, these constraints are insufficient to align latent spaces globally, as none of the introduced loss functions have locally bounded properties. Yet, another constraint that is not restricted to local properties is necessary.

In Riemannian geometry, it is common to attack geometric equations to find a specific form of a metric of the given space. If one can find the explicit form of a metric, then the curvature of a given space can be identified, which can be utilized to understand the global characteristics of the space. Or, in other way around, if one has distance among points on a manifold, it is possible to find a metric from a distance equation.

$$S^2 = \int_l \sum_\mu \sum_\nu g_{\mu\nu} dx^\mu dx^\nu \tag{12}$$

However, in general, finding the analytic form of the metric is extremely complicated (or impossible). Therefore, we propose an idea to bypass this issue by utilizing the general mathematical characteristic of Riemannian geometry. In a curved space, distances between points are not intuitive and simple to compute. The metric is necessary to find finite distances. However, there is a wonderful invariance known as diffeomorphism in Riemannian manifolds. This invariance guarantees the freedom to fix coordinates by transformations induced by the Jacobian of a vector. And it is well-known that a locally flat frame is always possible to find around a given vector on a manifold. The locally flat frame, by its nature, is flat around the infinitesimal boundary of a vector. Therefore, the distance equation can now be reduced to a simpler form in local boundaries.

$$\begin{aligned} S^2 &= \int_l \sum_\mu \sum_\nu g_{\mu\nu} dx^\mu dx^\nu \\ &= \int_l \sum_\mu \sum_\nu \eta_{\mu\nu} dx^\mu dx^\nu \\ &= \int_a^b dx^2 \end{aligned} \tag{13}$$

Here, a indicates a given latent vector and b is a perturbation around vector a. If this perturbation is infinitesimal, the distance between the vector and its perturbation can be simplified as follows.

$$S = |b - a| \tag{14}$$

Now, for a given SMILES input and its infinitesimal perturbations, the latent vectors from the source and target tasks can be transformed into a vector on a universal manifold where the locally flat frame resides. One can compute distances between the latent vector and its perturbations from each task and require them to be the same. By doing so, the locally flat latent spaces will align together on a universal manifold and cover the overlapping region smoothly. Then, the mutual information can naturally be transferred from one to another, and the extrapolation performance of the model will be boosted by source data. In an abstract form, the distance loss can be expressed as follows.

$$l_{dis} = \frac{1}{M} \sum_\alpha C_\alpha \sum_i^M \mathrm{MSE}(s_\alpha^i, s_t^i) \tag{15}$$

Where M is the number of perturbations, C_α is the given distance ratio for source to target, and s_α^i is the displacement between pivot data points and their perturbations.

$$s_\alpha^i \equiv |(z'_\alpha) - (z'^i_\alpha)| \qquad s_t^i \equiv |(z'_t) - (z'^i_t)| \tag{16}$$

$$z'^i_\alpha = \mathrm{Transfer}_{\alpha \to LF}(\mathrm{Encoder}_\alpha(x^i)) \tag{17}$$

$$z'^i_t = \mathrm{Transfer}_{t \to LF}(\mathrm{Encoder}_t(x^i)) \tag{18}$$

Here x^i denotes ith perturbation of embedded input x, and $\mathrm{Encoder}_\alpha$ and $\mathrm{Encoder}_t$ are encoder parts of the source and target model, respectively. Finally, by gathering all losses with individual hyperparameters, we define the complete form of the loss function used in the extended *GATE* algorithm.

$$l_{tot} = l_{reg} + \alpha l_{auto} + \beta l_{cons} + \gamma l_{map} + \delta l_{dis} \tag{19}$$

Table 1. Regression performance of 10-task *GATE*, MTL, and STL in Pearson correlation.

Tasks	GATE	MTL	STL
Parachor	0.9309±0.0073	**0.9358±0.0060**	0.9287±0.0086
Surface Tension	**0.8440±0.0073**	0.8195±0.0236	0.7171±0.0211
Dielectric Constant	**0.9228±0.0169**	0.9099±0.0176	0.9216±0.0070
Hydration Free Energy	**0.9504±0.0107**	0.9409±0.0160	0.9414±0.0097
Heat of Vaporization	0.8962±0.0057	**0.9018±0.0067**	0.8618±0.0160
Boiling Point	**0.9113±0.0066**	0.9076±0.0087	0.8847±0.0316
Refractive Index	**0.9793±0.0025**	0.9781±0.0030	0.9761±0.0009
Density	**0.8581±0.0115**	0.8512±0.0143	0.8237±0.0330
Melting Point	0.8739±0.0052	0.8714±0.0073	**0.8901±0.0019**
Viscosity	**0.9105±0.0072**	0.8952±0.0061	0.8967±0.0134
No. 1st	**7**	2	1
Avg. Rank	**1.3**	2.2	2.5

Hyperparameters play a crucial role in weighted summation parameters, and by tuning them sophisticatedly, the model's performance will reach its peak. In most cases, many hyperparameters are sufficient to be set to a trivial number like 1, but for parameters γ, δ, and C_α, it is worthwhile to tune them for optimal model performance. However, finding the right combinations of parameters can be challenging due to the immense search space. In such cases, we can rely on scientific knowledge to guide us in tuning them.

3. Experiments

3.1. Experimental Setup

A total of 10 datasets curated from five different sources named PubChem(Kim et al., 2022), Ochem(Sushko et al., 2011), CCDDS, Yaws Handbook, and Jean-Claude Bradley were used for these experiments. We prepared the training and test sets by splitting each dataset according to the scaffold of the molecular structure(Bemis & Murcko, 1996). A single NVIDIA A40 was used for every experiment, and four-fold cross-validation setting with uniform sampling and a separate test set was used for the default setup. We used the same model architecture and hyperparameters for GAM model as described in Ko et al. (2023b). In all experiments, the encoder and head architecture were identical for GAM, MTL, and STL.

3.2. Effect of multi-task extension from two-task *GATE* to three-task *GATE*

We first compared the regression performance of three-task *GATE* and two-task *GATE* to assess the impact of multi-task extension. In each experiment, we used refractive index and heat of vaporization as pivot tasks and selected an additional task to constitute three tasks. Overall three sets of experiments were performed using hydration free energy, surface tension or boiling point as an additional tasks respectively. To assess the regression performance of the two-task *GATE*, we separately trained and averaged all three possible combinations of the three tasks.

As depicted in Figure 3, the results demonstrate a clear synergy effect among the three tasks. Across all three experiment sets, there is a consistent reduction in the root mean square error (RMSE) of the three-task *GATE* compared to the two-task *GATE*, even when different additional tasks are included in the sets. This result indicates that the prediction performance of molecular properties can be enhanced by incorporating suitable auxiliary tasks, and this synergy effect can be achieved through the proposed multi-task extension of the *GATE*.

3.3. Regression performance of many-task *GATE*

To assess the effectiveness of *GATE* for multi-task learning, we also compared the regression performance of the many-task *GATE* with that of classical multi-task learning (MTL) techniques and single task learning (STL). As shown in Table 1, Pearson correlation of *GATE* outperforms MTL and STL for 7 out of 10 tasks, whereas MTL and STL perform best for only 2 tasks and 1 task respectively. Moreover, GATE's regression performance ranks within the top 2 for all tasks, demonstrating robust performance with an average rank of 1.3.

The robustness of the *GATE* for many-task setup is even more clearly shown in Table 2. Table 2 presents percentage of improvement on regression performance of *GATE* and MTL compared to STL. As shown in the table, in many cases, multi-task setup enhances regression performance, but in some cases, it can actually reduce regression performance. This decline in performance can be attributed to the negative transfer of undesired interfering information among the tasks. As evident from the table, *GATE* shows a

reduction of performance in only one task, while classical MTL exhibits a performance decrease in four tasks out of ten tasks.

Table 2. Relative improvement of the regression performance of 10-task GATE and MTL over STL in percent.

Tasks	GATE	MTL
Parachor	0.24	0.78
Surface Tension	17.69	14.28
Dielectric Constant	0.13	**-1.26**
Hydration Free Energy	0.96	**-0.06**
Heat of Vaporization	3.99	4.64
Boiling Point	3.01	2.59
Refractive Index	0.32	0.21
Density	4.18	3.33
Melting Point	**-1.83**	**-2.09**
Viscosity	1.54	**-0.16**

The result is well aligned with the experiments on stability of the latent spaces introduced in the original *GATE* paper(Ko et al., 2023b), which showed that the latent space of GATE exhibits relatively stable characteristics compared to that of MTL. Because the *GATE* is more resilient to interfering information, it exhibits more robust regression performance in a multi-task setup involving numerous tasks, where there is complex information exchange among the tasks.

4. Discussion

The original GATE algorithm interprets the latent space as a curved space and utilizes the mathematical concept of differential geometry, particularly Riemannian manifolds. Since the mathematical concept of GATE is not restricted to the two-task case, it is straightforward to generalize the algorithm to cover multiple source tasks without loss of generality. In this work, we designed the mathematical notion of the extended GATE with newly introduced hyperparameters and extended losses, and we have demonstrated the superior performance of the model using numerous open database datasets.

While our model outperforms conventional setups, there are several areas for improvement. First, the model's computational complexity grows significantly with the number of source tasks. Since the distance and mapping losses must be computed for every pair of source and target tasks, the complexity is on the order of $\mathcal{O}(N^2)$. Therefore, compactifying the model architecture is one research direction to explore.

Second, the distance loss can potentially be omitted if one can directly calculate the curvature of the space by finding the analytic form of the metric tensor. While this is normally impossible, by utilizing the notion of operator learning, it can be achieved. After specifying the form of the metric tensor, one can pre-calculate the Ricci scalar of the space in advance. By matching the Ricci scalar from source and target spaces, the distance loss can be omitted and replaced. This idea can encode geometric information not restricted to local geometry but global, potentially improving GATE's performance and robustness even further.

References

Stephan Allenspach, Jan A Hiss, and Gisbert Schneider. Neural multi-task learning in drug design. *Nature Machine Intelligence*, 6(2):124–137, 2024.

Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs. 12 2013.

Rich Caruana. Multitask learning. *Machine learning*, 28: 41–75, 1997.

Connor W. Coley, Wengong Jin, Luke Rogers, Timothy F. Jamison, Tommi S. Jaakkola, William H. Green, Regina Barzilay, and Klavs F. Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.*, 10:370–377, 2019. doi: 10.1039/C8SC04228D. URL http://dx.doi.org/10.1039/C8SC04228D.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. 06 2016.

David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in Neural Information Processing Systems (NIPS)*, 13, 09 2015.

Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi Jaakkola. Learning multimodal graph-to-graph translation for molecular optimization, 12 2018.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 10 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac956. URL https://doi.org/10.1093/nar/gkac956.

Sung Moon Ko, Sungjun Cho, Dae-Woong Jeong, Sehui Han, Moontae Lee, and Honglak Lee. Grouping matrix based graph pooling with adaptive number of clusters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8334–8342, 2023a.

Sung Moon Ko, Sumin Lee, Dae-Woong Jeong, Woohyung Lim, and Sehui Han. Geometrically aligned transfer encoder for inductive transfer in regression tasks, 2023b.

Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. *CVPR 2011*, pp. 1785–1792, 2011. URL https://api.semanticscholar.org/CorpusID:7419723.

Shengchao Liu, Meng Qu, Zuobai Zhang, Huiyu Cai, and Jian Tang. Structured multi-task learning for molecular property prediction. In *International conference on artificial intelligence and statistics*, pp. 8906–8920. PMLR, 2022.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Wei Cheng, Xiang Zhang, and Wei Wang. *Dual Transfer Learning*, pp. 540–551. doi: 10.1137/1.9781611972825.47. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611972825.47.

Jianhan Pan, Teng Cui, Thuc Duy Le, Xiaomei Li, and Jing Zhang. Multi-group transfer learning on multiple latent spaces for text classification. *IEEE Access*, 8:64120–64130, 2020. doi: 10.1109/ACCESS.2020.2984571.

Minshi Peng, Yue Li, Brie Wamsley, Yuting Wei, and Kathryn Roeder. Integration and transfer learning of single-cell transcriptomes via cfit. *Proceedings of the National Academy of Sciences*, 118(10):e2024383118, 2021. doi: 10.1073/pnas.2024383118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2024383118.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. Transfer learning for image classification with sparse prototype representations. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2, 03 2008. doi: 10.1109/CVPR.2008.4587637.

Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning with applications to medical imaging. *CoRR*, abs/1902.07208, 2019. URL http://arxiv.org/abs/1902.07208.

Franco Scarselli, Marco Gori, Ah Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 20:61–80, 01 2009. doi: 10.1109/TNN.2008.2005605.

Iurii Sushko, Sergii Novotarskyi, Robert Körner, Anil Kumar Pandey, Matthias Rupp, Wolfram Teetz, Stefan Brandmaier, Ahmed Abdelaziz, Volodymyr V Prokopenko, Vsevolod Y Tanchuk, et al. Online chemical modeling environment (ochem): web platform for data storage, model development and publishing of chemical information. *Journal of computer-aided molecular design*, 25:533–554, 2011.

Jingshu Wang, Divyansh Agarwal, Mo Huang, Gang Hu, Zilu Zhou, Chengzhong Ye, and Nancy Zhang. Data denoising with transfer learning in single-cell transcriptomics. *Nature Methods*, 16:875–878, 09 2019. doi: 10.1038/s41592-019-0537-1.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Tim Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59, 07 2019. doi: 10.1021/acs.jcim.9b00237.

Xiang Yu, Jian Wang, Qing-Qi Hong, Raja Teku, Shui-Hua Wang, and Yu-Dong Zhang. Transfer learning for medical images analyses: A survey. *Neurocomputing*, 489:230–254, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.08.159. URL https://www.sciencedirect.com/science/article/pii/S0925231222003174.

Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

Fuzhen Zhuang, Ping Luo, Hui Xiong, Qing He, Yuhong Xiong, and Zhongzhi Shi. Exploiting associations between word clusters and document classes for cross-domain text categorization†. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4 (1):100–114, 2011. doi: https://doi.org/10.1002/sam.10099. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.10099.

Fuzhen Zhuang, Ping Luo, Changying Du, Qing He, and Zhongzhi Shi. Triplex transfer learning: Exploiting both shared and distinct concepts for text classification. In *Proceedings of the Sixth ACM International*

Conference on Web Search and Data Mining, WSDM '13, pp. 425–434, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318693. doi: 10.1145/2433396.2433449. URL https://doi.org/10.1145/2433396.2433449.

Fuzhen Zhuang, Ping Luo, Changying Du, Qing He, Zhongzhi Shi, and Hui Xiong. Triplex transfer learning: Exploiting both shared and distinct concepts for text classification. *IEEE Transactions on Cybernetics*, 44(7): 1191–1203, 2014. doi: 10.1109/TCYB.2013.2281451.