Multi-modal Foundation Model for Material Design

Seiji Takeda
IBM Research - Tokyo

Indra Priyadarsini
IBM Research - Tokyo

Akihiro Kishimoto
IBM Research - Tokyo

Hajime Shinohara
IBM Research - Tokyo

Lisa Hamada
IBM Research - Tokyo

Masataka Hirose
JSR Corporation

Junta Fuchiwaki
JSR Corporation

Daiju Nakano
IBM Research - Tokyo

Abstract

We propose a multi-modal foundation model for small molecules, a shift from traditional AI models that are tailored for individual tasks and modalities. This model uses a late fusion strategy to align and fuse three distinct modalities: SELFIES, DFT properties, and optical spectrum. The model is pre-trained with over 6 billion samples to provide two primary functions, generating fused feature representations across the three modalities, and cross-modal predictions and genrations. As preliminary experiments, we demonstrate that the fused representation successfully improves the performance of property predictions for chromophore molecules, and showcase 6 distinct cross-modal inferences.

1 Introduction

Deep-learning models have been playing key roles in artificial intelligence (AI)-guided molecular design. A wide variety of models have been reported to carry out numerous tasks including property prediction (1; 2; 3), structure generation (4; 5; 6) and chemical-synthesis prediction (7; 8; 9) for small organic molecular design.

However, since these models have been independently developed by individual researchers or groups, most are isolated and siloed in terms of the following three aspects. (i) Data modality: each model handles only a single data modality corresponding a specific molecular representation such as SMILES (10; 11), even though a molecule can be represented by a wide variety of data modalities. (ii) Material domain: each model is trained with only the specific material domain's data such as OLED (12; 13). (iii) Application task: each model is trained to address only specific tasks, e.g. prediction of properties such as melting temperature (14; 15). The above-mentioned inter-model dislinkages result in several drawbacks. One is that knowledge captured by each model is not shared with other models, being different from chemists' way of thinking to overarch different data and domains. Another is only limited amount of data addressing specific aspects (modality and task) can be used to train each model. These, thus, have a negative impact on the accuracy of the feature representations captured with the models, hence on prediction and generation tasks.

In contrast to these discrete traditional AI models, a foundation model (FM) is a more comprehensive model pre-trained with a large training data set in a self-supervised manner and developed to capture foundational representations (16). It is adaptable to a variety of downstream tasks (e.g., language translation, and summary) with a little additional training. BERT (17), GPT-3 (18), and DALL-E (19) are well-known examples. Recently multi-modal FMs which use data from various modalities (e.g., text, video, and speech), have been proposed (20; 21; 22; 23; 24). Such models, by aligning or

37th Conference on Neural Information Processing Systems (NeurIPS 2023) AI4MAT Workshop.

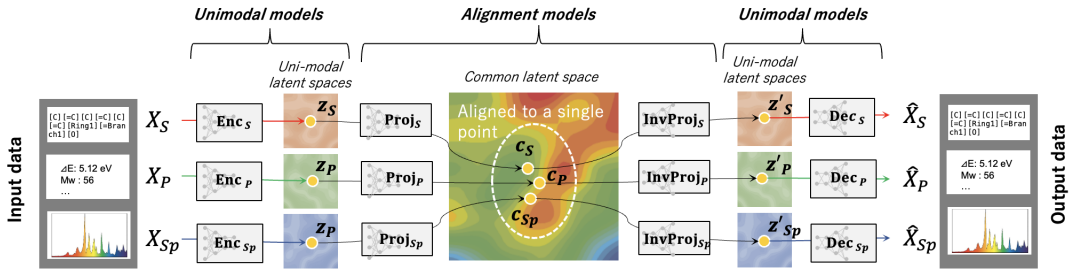

Figure 1: Architecture of our multi-modal FM for material design

fusing multiple input modalities in a latent space, achieve downstream tasks with higher precision than single-modal models. Similarly, it can be speculated that a multi-modal FM for materials can effectively capture the material representations and improve the predictions in downstream tasks. However, availability of large datasets is a major constraint in developing a multi-modal FM for material domain (25). The amount of dataset available for any two modalities vary in several orders of magnitude. For example, the largest dataset available for SMILES is about 37B while there are only a few thousands of UV spectrum dataset. Therefore, it is crucial to design a suitable model architecture that takes into account the data sparsity and missing modalities.

We propose a multi-modal FM that is trained with massive sparse data and captures comprehensive representations of molecules, so that the model can be applied to a wide variety of downstream tasks.

2 Method

In contrast to training a single massive multi-modal model, we propose a FM with a distinct architecture and learning strategy as illustrated in Figure 1. The proposed FM comprises of two types of models: (1) *uni-modal model*, an encoder-decoder based model trained exclusively for each modality; and (2) *alignment model*, made up of projector-inverse projector networks that is trained to project and align the unimodal latent representations onto a common latent space and vice versa. This strategy offers two main advantages : (1) it can handle data samples with missing modalities and (2) it provides a modular structure and flexibility to add other unimodal models to the FM, requiring only a minimal tuning of the alignment models.

In this paper, we consider three modalities; SELFIES (SELF-referencing Embedded Strings), Density Functional Theory (DFT) Properties, and Spectrum. The input dataset for each of the modalities SELFIES (S), Properties (P), and Spectrum (Sp) are denoted as $\mathbf{X_S}$, $\mathbf{X_P}$, and $\mathbf{X_{Sp}}$.

The uni-modal models are pre-trained in a self-supervised manner on their respective modality datasets such that for each uni-modal model, given the input data $\mathbf{X_i}$, the encoder gives the corresponding uni-modal latent representation $\mathbf{z}_i$,

$$\mathbf{z}_i = \text{Encoder}_i(\mathbf{X}_i), \qquad \text{where } i \in \{S, P, Sp\}$$

and from $\mathbf{z}_i$, the decoder reconstructs the original input. This pre-training is independently performed until the encoder-decoder of each uni-modal model is sufficiently trained.

Next, in order to capture the combined representation of all the modalities, we introduce the alignment models between the unimodal encoder-decoder pair. The uni-modal latent representations $\mathbf{z}_{i\in\{S,P,Sp\}}$ are projected onto a common latent space through projectors $Proj_{i\in\{S,P,Sp\}}$,

$$\mathbf{c}_i = \text{Proj}_i(\mathbf{z}_i), \qquad \text{where } i \in \{S, P, Sp\}$$

The projectors are trained to minimize the distance L between the projected vectors $\mathbf{c_{i\in\{S,P,Sp\}}}$.

$$\min L = \sum_{i,j\in\{S,P,Sp\},i\neq j} \|\mathbf{c}_i - \mathbf{c}_j\|^2$$

It should be noted that most samples include missing modalities. When the full set of modalities $\{S, P, Sp\}$ is not available, the projectors are trained sequentially for each available pair of modalities.

An inverse-projector $InvProj_{i \in \{S,P,Sp\}}$ for each modality is then trained to reconstruct the respective original uni-modal latent representations from the aligned common representation.

$$\mathbf{z}'_i = \text{InvProj}_i(\mathbf{c}_i), \quad \text{where } i \in \{S, P, Sp\}$$

$$\min L_{\text{inv}} = \sum_{i \in \{S,P,Sp\}} \|\mathbf{z}'_i - \mathbf{z}_i\|^2$$

Finally, the uni-modal model decoder are used to reconstruct the original input from the inverse-projected uni-modal representations.

$$\hat{\mathbf{X}}_i = \text{Decoder}_i(\mathbf{z}'_i), \quad \text{where } i \in \{S, P, Sp\}$$

Armed with the common latent space configuration, our FM delivers two primary functionalities in the inference phase. The first is providing fused representations; combining the aligned representation with uni-modal representations. The fused feature captures wider aspects of molecules so that it can improve downstream tasks such as property preditions. The second functionality is cross-modal prediction and generation. For example, given input SELFIES (molecular structure), our FM can predict the corresponding properties and generate a spectrum form as output, and vice versa.

3 Experimental Setup

In the pre-training phase, each uni-modal model, projectors, and inverse projectors are sequentially trained by corresponding data sets. The details are described below.

The uni-modal model for SELFIES (26) is a transformer model based on the BART architecture (27). Comparing with SMILES (28; 29; 30), SELFIES provides a more concise representation, the grammar of which is robust (26; 31; 32). The data set consists of 6 billion molecular structures from the ZINC22 dataset (33). Each SELFIES string is tokenized by a wordlevel tokenizer and applied with 15% masks. The BART model is implemented with two encoder and decoder layers, four attention heads, and six hidden layers.

The Properties model is configured with a variational autoencoder (VAE) with fully connected layers. DFT (Density Functional Theory)-calculated properties extracted from QM9 data set (34), containing 134,000 samples, is used. We use 12 properties which are μ: dipole moment, α: isotropic polarizability, E_{homo}: HOMO energy, E_{lumo}: LUMO energy, E_{gap}: energy gap, r^2: electronic spatial extent, $zpve$: zero-point vibrational energy, U_0: internal energy at 0 K, U: internal energy at 298 K, H: enthalpy, G: free energy, and C_V: heat capacity.

The Spectrum model is configured with a VAE with gated recurrent units (GRUs). The training data set contains about 1,000 samples, each of which consists of 180 data points of optical absorption spectrum plots used in a previous study (35).

After pre-training the uni-modal models, the projectors and inverse projectors for each of these modalities are trained. It must be noted that there are few publicly available datasets for all the three modalities. Hence we undertake a pairwise training strategy where first, any two modalities are aligned onto a common latent space by training their projectors. Next, the third modality is aligned with the first one. Thus adding other modalities can also be carried out as a post hoc process. The projectors for SELFIES and Properties are trained using the 134,000 SELFIES-Properties samples from the QM9 dataset. The projector for Spectrum is then trained with the 1000 SELFIES-Spectrum samples from the spectrum dataset. Lastly, the inverse projectors are trained to reconstruct the original unimodal latent spaces. The entire training process required about 300 hours with 8 A100 GPUs.

4 Results and Discussion

4.1 Downstream Task1: Fused representation

We evaluated the latent representations obtained from the FM to train a downstream predictive model. To train the predictive model, we used experimental dataset including approximately 17,000 samples, consisting of SMILES of chromophore molecules' solute and solvent, and the corresponding peak emission wavelengths (36). These samples were divided into training and test sets with the ratio

Table 1: Prediction results of different representations and downstream models

	Kernel-Ridge		DNN	
	R^2	RMSE (nm)	R^2	RMSE (nm)
FM and MFP	**0.931**	**27.6**	0.911	30.5
FM	0.912	30.1	0.882	36.5
MFP	0.906	32.5	0.878	39.2

of 0.8 and 0.2, respectively. For the downstream prediction, we employed kernel ridge and fully connected deep neural network (DNN) regression models. Details of the models are illustrated in Supplementary A. Only molecular structure (SELFIES) is given as input to the FM, and its fused representation is provided as a concatenation of the aligned representation $\mathbf{c}_{\mathrm{S}}$ and the uni-modal SELFIES representation $\mathbf{z}_{\mathrm{S}}$. We also computed a widely used molecular descriptor Morgan Fingerprint $\mathbf{x}_{\mathrm{MFP}}$. We then evaluated the fusion of these representations with independent ones.

The results of test set over different representations and downstream models to predict peak wavelengths are listed in Table 1 and the corresponding plots are illustrated in Supplementary B. It is confirmed that the representation captured by our model exceeded the accuracy of the conventional fingerprint, indicating the supremacy of large data-driven features encapsulating various aspects of molecules (e.g., DFT properties). Furthermore, the fusion of those features achieved the highest prediction accuracy, implying that our FM representation works complementarily with other features. It is promising to enhance the predictive performance with the inclusion of other representations.

4.2 Downstream Task2: Cross-modal inference

Alignment of the latent representations of each modality within a common latent space enabled the prediction and generation of multi-modal outputs from one input modality. Example results of the cross-modal inferences for various modalities are presented in Figure 2. In the generative inference, due to the nature of the BART decoder, molecular structures, as represented by SELFIES, were generated with a myriad of variations by accepting broader prioritized output tokens.

The advantage is that six distinct predictive and generative inferences are achieved by a single FM without accessing independent task-specific models. Due to the split pre-training strategy, a user can integrate other modalities with post hoc training of only the alignment models. This enables $\binom{N}{2}$ possible inference tasks by the single FM given that N modalities are integrated.

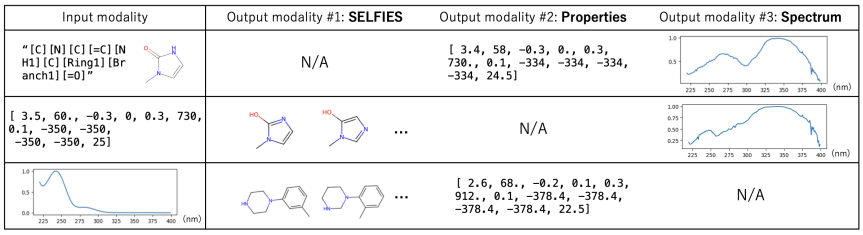

Figure 2: Example results of cross-modal inferences. Molecular images are depicted instead of SELFIES. Inferences within the same modalities are denoted as N/A.

4.3 Limitations and Future Work

One current limitation is that the molecular structures associated with DFT properties and spectra are considerably distinct. This forces the cross-modal inference for these modalities a significant extrapolation. To overcome this, we are collecting molecular structures across a wider range of modalities from sources such as PubChem and patents. Another limitation is that available data set is small for alignment training, because we use only paired data. To overcome this, we leverage additional alignment strategy such as contrastive learning, where non-paired data can be used too. Finally, we are expanding our model to billion parameters and integrating additional modalities (e.g. 3D atom coordinates). These ongoing efforts promise to further enhance our model's capabilities.

5 Conclusion

We introduced a multi-modal foundation model for material design pre-trained with over 6 billion samples. The architecture harmoniously aligns three specific uni-modal models: BART for SELFIES, VAEs for Properties and Spectra, all within a common latent space. It is confirmed that a fused representation outperforms an existing representation. 6 distinct cross-modal inferences are illustrated too. These are promising results to scale our FM in both parameters and modalities.

References

[1] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, "Low data drug discovery with one-shot learning," *ACS Central Science*, vol. 3, no. 4, pp. 283–293, 2017.

[2] J. A. Fine, A. A. Rajasekar, K. P. Jethava, and G. Chopra, "Spectral deep learning for prediction and prospective validation of functional groups," *Chemical Science*, vol. 11, pp. 4618–4630, 2020.

[3] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams, and A. Aspuru-Guzik, "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach," *Nature Materials*, vol. 15, pp. 1120–1127, 2016.

[4] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *ACS Central Science*, vol. 4, no. 2, pp. 268–276, 2018.

[5] J. Lim, S.-Y. Hwang, S. Moon, S. Kimb, and W. Y. Kim, "Scaffold-based molecular design with a graph generative model," *Chemical Science*, vol. 11, pp. 1153–1164, 2020.

[6] S. Takeda, T. Hama, H.-H. Hsu, V. A. Piunova, D. Zubarev, D. P. Sanders, J. W. Pitera, M. Kogoh, T. Hongo, Y. Cheng, W. Bocanett, H. Nakashika, A. Fujita, Y. Tsuchiya, K. Hino, K. Yano, S. Hirose, H. Toda, Y. Orii, and D. Nakano:, "Molecular inverse-design platform for material industries," in *KDD*, pp. 2961–2969, 2020.

[7] A. Kishimoto, B. Buesser, B. Chen, and A. Botea, "Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning," in *NeurIPS*, pp. 7224–7234, 2019.

[8] J. S. Schreck, C. W. Coley, and K. J. M. Bishop, "Learning retrosynthetic planning through simulated experience," *ACS Central Science*, vol. 5, no. 6, pp. 970–981, 2019.

[9] M. H. S. Segler, M. Preuss, and M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic AI," *Nature*, vol. 555, pp. 604–610, 2018.

[10] B. Sattarov, I. I. Baskin, D. Horvath, G. Marcou, E. J. Bjerrum, and A. Varnek, "De novo molecular design by combining deep autoencoder recurrent neural networks with generative topographic mapping," *Journal of chemical information and modeling*, vol. 59, no. 3, pp. 1182–1196, 2019.

[11] N. De Cao and T. Kipf, "Molgan: An implicit generative model for small molecular graphs. arxiv 2018," *arXiv preprint arXiv:1805.11973*, 2019.

[12] H. Abroshan, P. Winget, H. S. Kwak, Y. An, C. T. Brown, and M. D. Halls, "Machine learning for the design of novel oled materials," in *Machine Learning in Materials Informatics: Methods and Applications*, pp. 33–49, ACS Publications, 2022.

[13] K. Kaufmann, D. Maryanovsky, W. M. Mellor, C. Zhu, A. S. Rosengarten, T. J. Harrington, C. Oses, C. Toher, S. Curtarolo, and K. S. Vecchio, "Discovery of high-entropy ceramics via machine learning," *Npj Computational Materials*, vol. 6, no. 1, p. 42, 2020.

[14] M. Hirohara, Y. Saito, Y. Koda, K. Sato, and Y. Sakakibara, "Convolutional neural network based on smiles representation of compounds for detecting chemical motif," *BMC bioinformatics*, vol. 19, pp. 83–94, 2018.

[15] G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. Da Silva, and M. G. Quiles, "Machine learning prediction of nine molecular properties based on the smiles representation of the qm9 quantum-chemistry dataset," *The Journal of Physical Chemistry A*, vol. 124, no. 47, pp. 9854–9866, 2020.

[16] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, and S. A. *et al.*, "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258, 2021.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *ACL*, pp. 4171–4186, 2019.

[18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners." arXiv preprint arXiv:2005.14165, 2020.

[19] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, pp. 8821–8831, 2021.

[20] W. Chen, M.-W. Chang, E. Schlinger, W. Wang, and W. W. Cohen, "Open question answering over tables and text," *arXiv preprint arXiv:2010.10439*, 2020.

[21] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7464–7473, 2019.

[22] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," in *NeurIPS*, 2021.

[23] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. D. Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-supervised multimodal versatile networks," in *NeurIPS*, pp. 25–37, 2020.

[24] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 1122–1131, 2020.

[25] D. Jha, V. Gupta, L. Ward, Z. Yang, C. Wolverton, I. Foster, W. k. Liao, A. Choudhary, and A. Agrawal, "Enabling deeper learning on big data for materials informatics applications," *Scientific Reports*, vol. 11, no. 4244, 2021.

[26] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, "Self-referencing embedded strings (selfies): A 100% robust molecular string representation," *Machine Learning: Science and Technology*, vol. 1, no. 4, p. 045024, 2020.

[27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[28] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: large-scale self-supervised pre-training for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[29] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, "Chemformer: a pre-trained transformer for computational chemistry," *Machine Learning: Science and Technology*, vol. 3, no. 1, p. 015022, 2022.

[30] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, "Large-scale chemical language representations capture molecular structure and properties," *Nature Machine Intelligence*, vol. 4, no. 12, pp. 1256–1264, 2022.

[31] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N. C. Frey, P. Friederich, T. Gaudin, A. A. Gayle, K. M. Jablonka, *et al.*, "Selfies and the future of molecular string representations," *Patterns*, vol. 3, no. 10, 2022.

[32] A. Yüksel, E. Ulusoy, A. Ünlü, and T. Doğan, "Selformer: Molecular representation learning via selfies language models," *Machine Learning: Science and Technology*, 2023.

[33] B. I. Tingle, K. G. Tang, M. Castanon, J. J. Gutierrez, M. Khurelbaatar, C. Dandarchuluun, Y. S. Moroz, and J. J. Irwin, "Zinc-22 - a free multi-billion-scale database of tangible compounds for ligand discovery," *Journal of Chemical Information and Modeling*, vol. 63, no. 4, pp. 1166–1176, 2023.

[34] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," *Scientific data*, vol. 1, no. 1, pp. 1–7, 2014.

[35] F. Urbina, K. Batra, K. J. Luebke, J. D. White, D. Matsiev, L. L. Olson, J. P. Malerich, M. A. Hupcey, P. B. Madrid, and S. Ekins, "Uv-advisor: attention-based recurrent neural networks to predict uv–vis spectra," *Analytical chemistry*, vol. 93, no. 48, pp. 16076–16085, 2021.

[36] J. F. Joung, M. Han, M. Jeong, and S. Park, "Experimental database of optical properties of organic compounds," *Scientific Data*, vol. 7, no. 295, 2020.

Supplementary

A. Downstream model configurations

One downstream predictive model, a Kernel-Ridge regression model, was configured using a RBF kernel with a hyperparameter α swept in the set $\{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0\}$.

Another downstream predictive model, a fully-connected DNN, was designed with various architectures having 4 layers. The node configurations are as follows:

$$D_{\text{in}} \to D_{\text{in}}/2 \to D_{\text{in}}/2 \to 1,$$
$$D_{\text{in}} \to D_{\text{in}}/2 \to D_{\text{in}}/4 \to 1,$$
$$D_{\text{in}} \to D_{\text{in}}/4 \to D_{\text{in}}/4 \to 1,$$

where D_{in} denotes the dimension of the input data (e.g., $z_S \oplus c_S$). In each model, drop rate was swept across 0.1, 0.2, 0.3, 0.4, 0.5. The training was performed with a learning rate of 0.001 over 1,000 epochs, in which the saturation of the test loss was confirmed.

The input dataset was tested with multiple configurations: with and without scaling, and with PCA (Principal Component Analysis) dimension reduction to dimension 32, 64, 128, 256, 512 as well as without PCA.

B. Downstream property prediction results

The plots of actual versus predicted peak wavelengths using the best downstream model is exhibited in 3. A successful fit for the majority of data points is confirmed. A few outliers will need investigation in future work using a larger model.

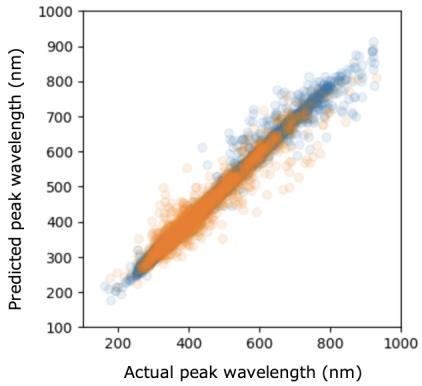

Figure 3: Peak wavelengths of actual and predicted values. Blue and orange plots represent training and test sets, respectively.