Trustworthy Inverse Molecular Design via Alignment with Molecular Dynamics

Kevin Tirta Wijaya [1] **Navid Ansari** [1] **Hans-Peter Seidel** [1] **Vahid Babaei** [1]

Abstract

Data-driven inverse molecular design (IMD) has attracted significant attention in recent years. Despite the remarkable progress, existing IMD methods lag behind in terms of *trustworthiness*, as indicated by their misalignment to the ground-truth function that models the molecular dynamics. Here, we propose TrustMol, an IMD method built to be trustworthy by inverting a reliable molecular property predictor. TrustMol first constructs a latent space with a novel variational autoencoder (VAE) and trains an ensemble of property predictors to learn the mapping from the latent space to the property space. The training samples for the ensemble are obtained from a new reacquisition method to ensure that the samples are representative of the latent space. To generate a desired molecule, TrustMol optimizes a latent design by minimizing both the predictive error and the uncertainty quantified by the ensemble. As a result, TrustMol achieves state-of-the-art performance in terms of IMD accuracy, and more importantly, aligned with the ground-truth function which indicates trustworthiness.

1. Introduction

Inverse molecular design (IMD) is a promising approach to accelerate the discovery of new molecules with desired properties. In IMD, molecules are designed to exhibit a target property, ideally by inverting the *native forward process* (NFP) (Ansari et al., 2022)—the ground-truth function that maps a molecule to their properties. However, such an inversion is extremely challenging. The common approach is to approximate NFP using a data-driven surrogate model.

Data-driven surrogate-based IMD approach has become increasingly popular, ranging from autoregressive models (Luo et al., 2021; Luo & Ji, 2022; Gebauer et al., 2019) to diffusion models (Hoogeboom et al., 2022; Xu et al., 2023). While prior works have progressively improved the state-of-the-art IMD accuracy, they have largely overlooked an equally critical aspect of IMD: *trustworthiness.*

The trustworthiness of a surrogate-based IMD method can be defined as how well it aligns with the NFP. For surrogate-based IMD methods that directly model the NFP (Gómez-Bombarelli et al., 2018; Eckmann et al., 2022), this alignment can be quantified by calculating the distance between surrogate-based error and the NFP-based error. For example, a surrogate might identify a molecule as a good match with a predicted property close to the desired property, resulting in a low surrogate error. But when the molecule is passed through the NFP, it proves to be a poor match (high NFP error) or invalid. An IMD method that lacks alignment with the NFP is not effective for discovering new molecules, as the NFP serves as the ground-truth representation of the molecule-to-property mapping in the real world.

In this work, we have identified two root issues to this misalignment problem: (**I1**) the surrogate fails to correctly model the forward process (e.g., the mapping from a molecular design space to the property space) on the training set, and (**I2**) the surrogate becomes unreliable when operating on molecules that are completely different from the training set, a scenario that often occurs during the inversion step.

We introduce TrustMol, a surrogate-based IMD method designed for trustworthiness by addressing the two issues. To improve forward modeling (**I1**), TrustMol uses a novel VAE to build a well-structured latent space and trains a surrogate to map latents directly to molecular properties. We also introduce a latent-property pairs *reacquisition* strategy that ensures the surrogate is trained on representative samples. During inference, TrustMol inverts the surrogate by optimizing a randomly-initialized molecular latent according to the distance between the predicted and the target property. To prevent the optimizer from exploring molecular latents far from the training data (**I2**), we incorporate *epistemic* uncertainty into the loss, guiding exploration toward regions where the surrogate is reliable.

We evaluate TrustMol against several state-of-the-art IMD baselines using two metrics: Mean Absolute Error (MAE) to the target property, and NFP-surrogate misalignment, which quantifies the gap between the property predictions

[1]Max Planck Institute for Informatics. Correspondence to: Kevin Tirta Wijaya <kwijaya@mpi-inf.mpg.de>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

of the surrogate and those of the NFP. TrustMol consistently outperforms the baselines across both metrics, achieving state-of-the-art performance in both single-objective and multi-objective IMD tasks. Source code is available at https://github.com/ktirta/TrustMol.

2. TrustMol

Given a desired property p, one way to implement the surrogate-based IMD approach involves using a property predictor surrogate Φ to optimize a randomly-initialized molecular design x,:

$$x^* = \arg\min_x |p - \Phi(x)|. \tag{1}$$

This approach is straightforward and has been used in multiple works (Gómez-Bombarelli et al., 2018; Eckmann et al., 2022). Unfortunately, its results are often misaligned with the NFP, i.e., molecules obtained are often deemed as poor matches by the NFP. From the forward modeling perspective, the mapping of molecular structures to their corresponding properties is inherently high-frequency, where small changes in structures can lead to significant changes in properties. This presents a challenge for neural networks, which tend to struggle to model high-frequency functions (Xu et al., 2019; Rahaman et al., 2019) (**I1**). Furthermore, not all molecular designs are valid; many molecular configurations are unstable and therefore invalid. From the inversion perspective, the optimization in equation (1) is unconstrained and can result in a molecular design that differ significantly from molecules in the training set. In this extrapolation regime, the surrogate prediction becomes unreliable (**I2**).

To address these challenges, we introduce three novel components: SGP-VAE (**I1**), latent-property pair reacquisition (**I1**), and uncertainty-aware molecular latent optimization (**I2**), described in detail in the following subsections.

2.1. Molecular Latent Optimization with SGP-VAE

We propose to perform the optimization in a well-structured latent space (Figure 1a (right) and 1b) to tackle the high-frequency and discontinuous nature of the molecule space. The latent space is learned by TrustMol through a VAE (Kingma & Welling, 2013) that is trained to reconstruct molecular representations from latent vectors. Our novel SELFIES-Graph-Property (SGP) VAE incorporates three sources of information, molecular strings, molecular 3D structures, and molecular properties information. Using SELFIES (Krenn et al., 2020) as the primary representation ensures that any latent vector can be decoded into a valid molecule. However, similarities between molecular strings are not highly correlated to similar properties. Therefore, we augment the VAE training with two auxiliary tasks: predicting properties directly from the latent vectors and reconstructing 3D molecular graphs. Learning latent-to-property predictions can organize the latent space with respect to property values (Gómez-Bombarelli et al., 2018), while 3D structural information is a useful indicator of similarity in property space (Martin et al., 2002). With the three training objectives, our SGP-VAE can learn a latent space in which similar latents are more likely to correspond to molecules with similar properties. As a result of the smoother mapping, the quality of the forward modeling is improved.

2.2. Latent-Property Pairs Reacquisition

Learning the mapping from latent space to property space is challenging and often results in poor prediction performance (Eckmann et al., 2022). This phenomenon arises from how the latent-to-property surrogate is trained. Given a VAE encoder Ψ^{enc} that has been pretrained on a dataset $\mathbb{D} = \{(m_i, p_i^{\text{gt}})\}$ where m_i is the i-th molecule and p_i^{gt} is its corresponding property, the common approach to train a latent-to-property surrogate Φ parameterized by ϕ is,

$$\mathbb{Z} = \{\Psi^{\text{enc}}(m_i)\} = \{\boldsymbol{z}_i\}, \tag{2}$$

$$\phi^* = \arg\min_\phi |p_i^{\text{gt}} - \Phi(\boldsymbol{z}_i)|, \tag{3}$$

where $i = 1, ..., |\mathbb{D}|$, ϕ^* is the optimal parameter of Φ and $\boldsymbol{z}_i$ is the latent representation of m_i. A limitation to this approach is that there are molecules in $\mathbb{D}$ that cannot be well-represented by the latent vectors. Encoding such molecules with Ψ^{enc} will produce valid latent vectors, but decoding them back with the decoder Ψ^{dec} will result in incorrect molecules due to non-zero VAE reconstruction errors. Training Φ to predict the properties of the latents of these problematic molecules would result in an unreliable surrogate.

Here, we propose a latent-property pairs reacquisition method to collect representative training samples for the surrogate. Utilizing Ψ^{dec} alongside a conformer generator h (RDKit, (RDKit, 2023)) and the NFP f (Psi4 (Smith et al., 2020)), we generate the new dataset $\mathbb{D}_{\text{new}}$ of latent-property pairs for training the surrogate Φ according to the following steps. First, latent representations $\boldsymbol{z}$ are randomly sampled from a Gaussian $\mathcal{N}(\mu, \sigma)$ with mean μ and variance σ,

$$\mathbb{Z}_{\text{new}} = \{\boldsymbol{z}_{\text{new, i}} \mid \boldsymbol{z}_{\text{new,i}} \sim \mathcal{N}(0, 1)\}_{i=1}^N. \tag{4}$$

The properties of the molecules represented by the sampled latents are then calculated by decoding the latents back into molecules using Ψ^{dec}, generating the corresponding 3D conformations using h, and passing the conformations to f, before collecting the pairs into one training dataset,

$$\mathbb{P}_{\text{new}} = \{f(h(\Psi^{\text{dec}}(\boldsymbol{z}_{\text{new, i}}))) \mid \forall \boldsymbol{z}_{\text{new, i}} \in \mathbb{Z}_{\text{new}}\}, \tag{5}$$

$$\mathbb{D}_{\text{new}} = \{(\boldsymbol{z}_{\text{new, i}}, p_{\text{new, i}}) \mid \forall \boldsymbol{z}_{\text{new, i}} \in \mathbb{Z}_{\text{new}} \text{ and } \forall p_{\text{new, i}} \in \mathbb{P}_{\text{new}}\}. \tag{6}$$

2.3. Uncertainty-aware Molecular Latent Optimization

A neural network is most reliable when performing prediction on samples from regions that are well-represented during training. In TrustMol, we incorporate *epistemic* uncertainty into the molecular latent optimization (Ansari et al., 2022) to guide the optimization into reliable regions. Since epistemic uncertainty is a measure of training data sparsity, minimizing it is equivalent to guiding the optimization toward molecular latents that are novel, but not completely different from latents that are available during training.

We use the predictive disagreement between accurate and diverse neural networks (Lakshminarayanan et al., 2017) to quantify the epistemic uncertainty. Here, we the surrogate is an ensemble of n multilayer perceptrons (MLPs) with identical number of layers but different activation functions. The surrogate model is trained to fit the NFP, i.e., $\{\Phi_j \mid \Phi_j : \boldsymbol{z} \mapsto \hat{p}\}, j = 1, ..., n$. Given the mean prediction $\Phi^{\text{avg}}(\boldsymbol{z}) = \frac{1}{N}\sum_{j=1}^{N}\Phi_j(\boldsymbol{z})$, the epistemic uncertainty (U) can be defined as,

$$\mathrm{U}(\boldsymbol{z}) = \frac{1}{N}\sum_{j=1}^{N}(\Phi_j(\boldsymbol{z}))^2 - (\Phi^{\text{avg}}(\boldsymbol{z}))^2. \tag{7}$$

The final uncertainty-aware IMD process of TrustMol (Figure 1c) obtains the optimal molecular latent $\boldsymbol{z}^*$ through gradient descent,

$$\boldsymbol{z}^* = \arg\min_{\boldsymbol{z}} |\Phi^{\text{avg}}(\boldsymbol{z}) - p| + \mathrm{U}(\boldsymbol{z}). \tag{8}$$

3. Experimental Section

3.1. Single-Objective Inverse Molecular Design

In single-objective IMD, we set our property of interest as either HOMO, LUMO, or dipole moment, as these three properties can be calculated using the DFT as the NFP with relatively high accuracy (Faber et al., 2017; Matuszek & Reynisson, 2016). We define our *target* property values as a set of $n = 2000$ evenly-spaced values within a specified range $[a, b]$ that covers both property values present in and absent from the training dataset. We set the ranges to [-10, 0] for HOMO, [-4, 2] for LUMO, and [0, 4] for dipole moment. Each IMD method has a budget of $k = 10$ tries to generate a molecule for each target property value, and we retain only the molecule exhibiting the lowest absolute error. Due to compute limitation, we set $n = 20$ for JANUS and omit its novelty and uniqueness metrics to ensure fairness with other methods that generate significantly more molecules. When using 2,000 CPU threads on AMD EPYC 7702 processors, the DFT-based molecular property calculation of 20K (i.e., $n \cdot k$) molecules takes around 6 hours to complete.

We employ four metrics to evaluate the methods. The **NFP Error** is the MAE between the DFT-calculated properties of the generated molecules and the target properties. We use **novelty** and **uniqueness** to measure the diversity of the generated molecular designs, with novelty representing the number of designs not present in the QM9 dataset (Ramakrishnan et al., 2014), and uniqueness representing the number of unique designs generated. We measure **latency** in two ways: *single*, the time to generate one molecule individually, and *batch*, the total time to generate multiple molecules in parallel.

As shown in Table 1, TrustMol outperforms all methods by a substantial margin in all three target property categories. These results demonstrate that improving explainability through a neural surrogate-based latent optimization approach does not compromise IMD accuracy. All methods also display high novelty, indicating the effectiveness of both denoising and property prediction networks for discovering novel molecules. However, existing optimization-based IMD methods tend to produce identical molecules, as reflected by their uniqueness. In contrast, TrustMol attains a high score for uniqueness that is competitive with state-of-the-art diffusion model, GeoLDM. The high uniqueness score can be attributed to the improved surrogate model of TrustMol, which, due to the latent-property pairs reacquisition, has been trained on a more diverse set of latent vectors, enabling it to navigate toward more diverse latent solutions during optimization. Similar to other optimization-based approaches, TrustMol can generate molecules within reasonable time frame, especially when compared to GeoLDM in batch generation setup where the latency of TrustMol is two orders of magnitude smaller.

3.2. Multi-Objective Inverse Molecular Design

While single-objective IMD has been commonly used in previous studies (Hoogeboom et al., 2022; Xu et al., 2023), real-world applications often involves multi-objective IMD. Therefore, we provide an analysis of multi-objective IMD performance of TrustMol and other IMD methods. In this comparison, the IMD methods are tasked with generating molecular designs that simultaneously exhibit specific values of HOMO, LUMO, and dipole moment.

As shown in Table 2, simultaneously optimizing for multiple properties tends to reduce the accuracy of IMD methods. Nevertheless, TrustMol manages to minimize the deterioration of its IMD accuracy, significantly outperforming others in all property categories. The superior performance of TrustMol can be attributed to the synergy of our uncertainty-aware optimization and latent-property pairs reacquisition for training the surrogate model.

3.3. Measuring Surrogate-NFP Alignment

For a neural surrogate-based IMD method to be considered reliable, it should demonstrate a reasonable alignment be-

Table 1. Experimental results for single-objective IMD (HOMO, LUMO, or Dipole Moment) over three runs. We measure the NFP error in electronvolt (eV) and Debeye (D). We also report the novelty, uniqueness, and latency, where the batch-latency is evaluated for generating 2000 molecules in parallel. Bolded values indicate the best performance on the column.

Model	NFP Error			Nov.	Uni.	Latency (s)	
	H (eV)	L (eV)	D (D)	(%)	(%)	single	batch
JANUS (Nigam et al., 2022)	3.29	0.80	0.90	-	-	7113	-
GeoLDM (Xu et al., 2023)	$1.16_{\pm 0.03}$	$0.39_{\pm 0.02}$	$0.56_{\pm 0.03}$	81.06	**94.26**	8.67	1617
SELFIES LDM	$0.97_{\pm 0.01}$	$0.33_{\pm 0.04}$	$0.95_{\pm 0.02}$	82.28	48.20	0.64	**0.77**
MGCVAE (Lee & Min, 2022)	$1.65_{\pm 0.03}$	$0.30_{\pm 0.01}$	$0.44_{\pm 0.02}$	**90.17**	85.97	**0.33**	6.55
SELFIES VAE (Gómez-Bombarelli et al., 2018)	$3.75_{\pm 0.29}$	$1.99_{\pm 0.20}$	$4.98_{\pm 0.04}$	21.26	7.82	8.57	-
LIMO (Eckmann et al., 2022)	$1.23_{\pm 0.18}$	$0.35_{\pm 0.14}$	$0.59_{\pm 0.08}$	87.80	21.30	4.12	7.80
TrustMol (ours)	$\mathbf{0.95_{\pm 0.06}}$	$\mathbf{0.25_{\pm 0.01}}$	$\mathbf{0.40_{\pm 0.02}}$	87.70	88.0	7.62	11.53

Table 2. Multi-objective NFP errors for various models across HOMO (H), LUMO (L), and Dipole Moment (D). Lower values indicate better performance.

Model	H (eV)	L (eV)	D (D)
JANUS	2.46	1.33	1.07
LIMO	$0.85_{\pm 0.05}$	$1.02_{\pm 0.05}$	$1.17_{\pm 0.11}$
MGCVAE	$2.26_{\pm 0.02}$	$0.71_{\pm 0.01}$	$3.76_{\pm 0.01}$
SELFIES VAE	$3.26_{\pm 0.26}$	$1.70_{\pm 0.17}$	$1.96_{\pm 0.02}$
TrustMol (ours)	$\mathbf{0.62_{\pm 0.03}}$	$\mathbf{0.63_{\pm 0.02}}$	$\mathbf{0.79_{\pm 0.03}}$

Table 3. NFP-surrogate error misalignment between TrustMol and other models. Misalignment is defined as the absolute difference between the NFP error and the surrogate error. Note that some methods cannot predict the surrogate errors.

Model	H (eV)	L (eV)	D (D)
JANUS	3.32	1.11	1.56
LIMO	$1.01_{\pm 0.07}$	$0.54_{\pm 0.06}$	$1.36_{\pm 0.32}$
MGCVAE	-	-	-
SELFIES VAE	$3.75_{\pm 0.29}$	$1.99_{\pm 0.20}$	$4.98_{\pm 0.04}$
TrustMol (ours)	$\mathbf{0.89_{\pm 0.13}}$	$\mathbf{0.25_{\pm 0.01}}$	$\mathbf{0.40_{\pm 0.02}}$

tween its surrogate and the NFP. This alignment can be evaluated by comparing the IMD errors as predicted by the surrogate (surrogate error) and those calculated by the NFP (NFP error). In the unlikely event when a surrogate-based IMD method is perfectly aligned with the NFP, the gap between the NFP and surrogate errors, i.e., the NFP-surrogate misalignment, is zero.

Table 3 shows the NFP-surrogate misalignment of several IMD methods. We can see that the misalignments of other surrogate-based IMD methods are relatively high. On the other hand, TrustMol achieves lower NFP-surrogate misalignment across all three property categories. These results validate our hypothesis that incorporating epistemic uncertainty into the optimization process can effectively reduce the NFP-surrogate misalignment, resulting in a more trustworthy IMD method.

4. Conclusion

We introduced TrustMol, a molecular latent optimization method that focuses on aligning with the NFP for a trustworthy IMD. TrustMol not only demonstrates superior performance over existing IMD methods in accuracy, but also excels in trustworthiness, as indicated by the low disagreement with the NFP. The effectiveness of TrustMol, however, is limited by the expressiveness of the latent space and the reliability of the surrogate model. Therefore, improving the latent space construction and the surrogation is crucial for a highly performant IMD. A promising path toward this goal is to explore the latent space further with active learning (Settles, 2009). We note that our uncertainty-aware molecular latent optimization is closely related to Bayesian optimization (BO) (Frazier, 2018). However, TrustMol follows an offline model-based optimization approach (Trabucco et al., 2022) and does not assume access to the NFP during the optimization, whereas BO requires frequent back and forth with the NFP (i.e., density functional theory (DFT)).

References

Ansari, N., Seidel, H.-P., Vahidi Ferdowsi, N., and Babaei, V. Autoinverse: Uncertainty aware inversion of neural networks. *Advances in Neural Information Processing Systems*, 35:8675–8686, 2022.

Eckmann, P., Sun, K., Zhao, B., Feng, M., Gilson, M., and Yu, R. Limo: Latent inceptionism for targeted molecule generation. In *International Conference on Machine Learning*, pp. 5777–5792. PMLR, 2022.

Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., Vinyals, O., Kearnes, S., Riley, P. F., and Von Lilienfeld, O. A. Prediction errors of

molecular machine learning models lower than hybrid dft error. *Journal of chemical theory and computation*, 13 (11):5255–5264, 2017.

Frazier, P. I. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Gebauer, N., Gastegger, M., and Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lee, M. and Min, K. Mgcvae: multi-objective inverse design via molecular graph conditional variational autoencoder. *Journal of chemical information and modeling*, 62(12): 2943–2950, 2022.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Luo, S., Guan, J., Ma, J., and Peng, J. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.

Luo, Y. and Ji, S. An autoregressive flow model for 3d molecular geometry generation from scratch. In *International Conference on Learning Representations (ICLR)*, 2022.

Martin, Y. C., Kofron, J. L., and Traphagen, L. M. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358, 2002.

Matuszek, A. M. and Reynisson, J. Defining known drug space using dft. *Molecular informatics*, 35(2):46–53, 2016.

Nigam, A., Pollice, R., and Aspuru-Guzik, A. Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. *Digital Discovery*, 1(4): 390–404, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

RDKit. Rdkit: Open-source cheminformatics. https://www.rdkit.org. 2023.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Settles, B. Active learning literature survey. 2009.

Smith, D. G., Burns, L. A., Simmonett, A. C., Parrish, R. M., Schieber, M. C., Galvelis, R., Kraus, P., Kruse, H., Di Remigio, R., Alenaizan, A., et al. Psi4 1.4: Open-source software for high-throughput quantum chemistry. *The Journal of chemical physics*, 152(18), 2020.

Trabucco, B., Geng, X., Kumar, A., and Levine, S. Design-bench: Benchmarks for data-driven offline model-based optimization. In *International Conference on Machine Learning*, pp. 21658–21676. PMLR, 2022.

Xu, M., Powers, A. S., Dror, R. O., Ermon, S., and Leskovec, J. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y., and Ma, Z. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

A. Methods

A.1. Dataset and Molecular Properties

We use the QM9 dataset (Ramakrishnan et al., 2014) as our initial training dataset $\mathbb{D}$ for the SGP-VAE. QM9 is a quantum chemistry dataset that consists of around 130K small molecules. Each molecule is represented at atomic-level, i.e., atom types and their corresponding 3D coordinates. The molecules contains up to 9 heavy atoms (C, N, O, F), and up to 29 atoms when including the Hydrogens. QM9 also provides various molecular properties including dipole moment, isotropic polarizability, Highest Occupied Molecular Orbital (HOMO), Lowest Unoccupied Molecular Orbital (LUMO), thermal capacity, among others.

In our experiments, we use HOMO, LUMO, and dipole moment as the potential target properties of the inversion. The gap between HOMO and LUMO can be used to predict the stability of a compound. Dipole moment, on the other hand, is a measure of a molecule's polarity, which in turn can be used to predict various physical properties such as solubility in water and boiling point.

A.2. Implementation Details

We implement all neural networks with PyTorch (Paszke et al., 2019). AdamW optimizer (Loshchilov & Hutter, 2017) and cosine annealing learning rate scheduler (Loshchilov & Hutter, 2016) are used in the optimization process for all models. We train the SGP-VAE for 50 epochs and the ensemble surrogate model for 300 epochs, with a batch size of 32. To improve diversity of the ensemble surrogate model, at each iteration, a subnetwork Φ_i in the ensemble has a probability of only $q = 0.3$ to perform a gradient descent step. This is equivalent to independently training each subnetwork for 90 epochs with different random seeds.

We use RDKit (RDKit, 2023) and Psi4 (Smith et al., 2020) as the NFP, the ground truth functions that model the behavior of molecules in real-world. RDKit is an open-source cheminformatics and machine learning software that can perform analysis on chemical structures. We use RDKit to generate the molecular conformation, i.e., the spatial arrangement of atoms in a molecule, of the SELFIES strings generated by LIMO (Eckmann et al., 2022) and TrustMol. Psi4 is an open-source quantum chemistry software that is capable of accurately predicting the properties of a molecular conformation using DFT. We use Psi4 to calculate the HOMO, LUMO, and dipole moment values of molecular conformations generated by the IMD methods.

A.3. Loss Function of the SGP-VAE

Our SGP-VAE architecture features an encoder Ψ^{enc} that takes as inputs the multivew representations of a molecule, $\boldsymbol{x}_{\text{selfies}}$ and $\boldsymbol{x}_{\text{graph}}$. The graph representation is processed with a graph neural network (EGNN, (Satorras et al., 2021)) before being fused with features from the SELFIES representation into a latent vector $\boldsymbol{z}$. During training, the VAE's decoder Ψ^{dec} reconstructs both SELFIES and graph representations and predict the properties of the molecule directly from its latent. The loss is calculated as follows,

$$\begin{aligned}\mathcal{L} = &|p_{\boldsymbol{x}} - \hat{p}_{\boldsymbol{x}}| + ||\boldsymbol{x}_{\text{graph}} - \hat{\boldsymbol{x}}_{\text{graph}}||_2^2 + \text{CE}(\boldsymbol{x}_{\text{selfies}}, \hat{\boldsymbol{x}}_{\text{selfies}}) \\ &+ \text{KL}(\boldsymbol{z}||\mathcal{N}(0,1)),\end{aligned} \tag{9}$$

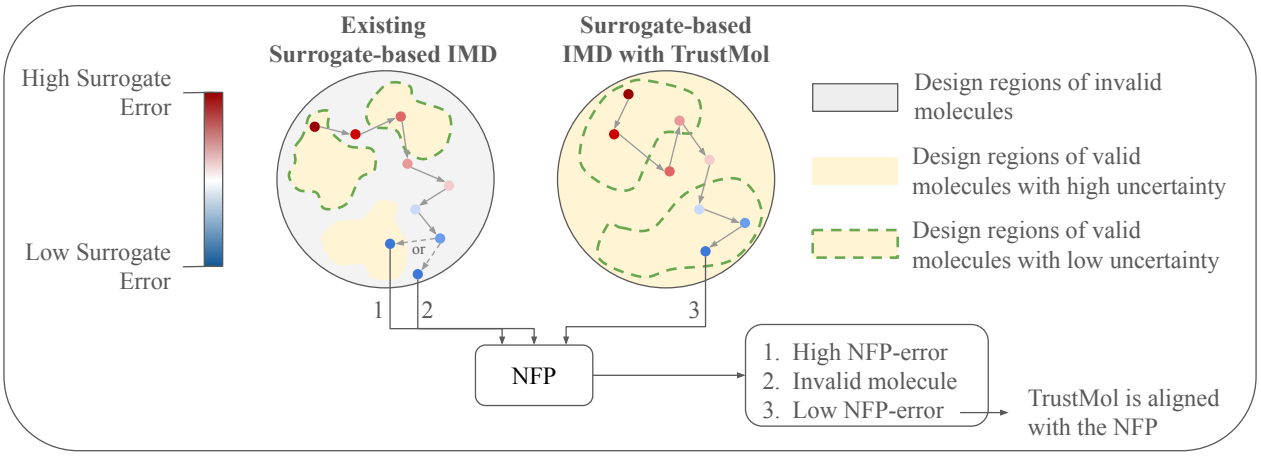
where CE and KL are cross-entropy and KL-divergence (Kullback & Leibler, 1951) loss functions, respectively.

B. Additional Results

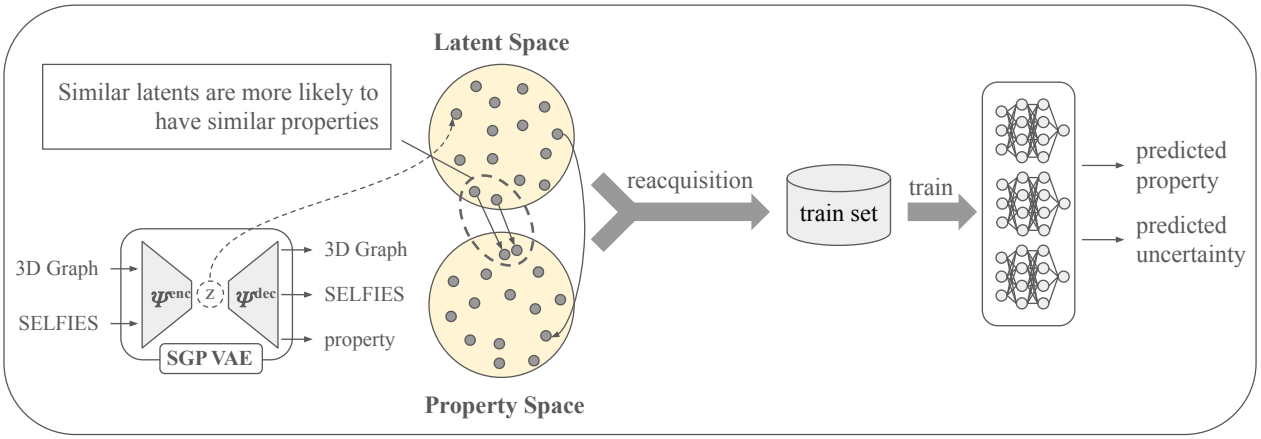
B.1. Verifying High-frequency and Discontinuous Nature of the Molecule Space

In earlier sections, we have discussed the high-frequency and discontinuous nature of the mapping from molecular space to property space, which has motivated us to choose molecular latents as our design representation. To validate our design choices, we analyze the impact of minimal noise injections on various molecular design representations with respect to their molecular properties.

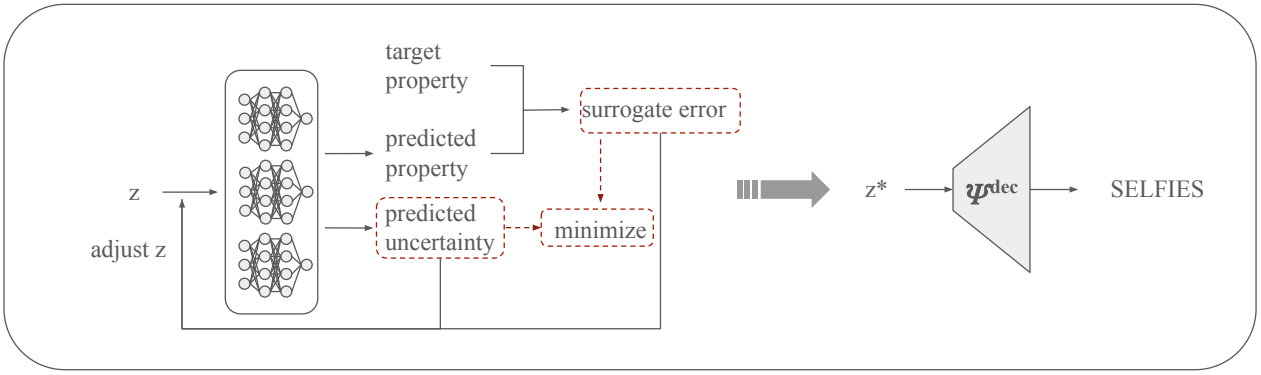
Table 4 shows the mean absolute error (MAE) between properties of the original and the noise-perturbed molecular designs. When noise from a $\mathcal{N}(0, 0.1)$ distribution is injected into a randomly-chosen atom coordinate of a 3D graph, the proportion



(a)



(b)



(c)

Figure 1. The framework of TrustMol. **(a)** Existing surrogate-based IMD often finds solutions in high-uncertainty regions that are far away from the training distribution, in which the surrogate predictions are most unreliable. This could lead to molecules that are invalid or have high NFP-error. TrustMol directs the IMD process into low-uncertainty regions where the surrogate can be trusted. **(b)** Improvement in the forward modeling comes from the SGP-VAE, which encourages similar latents to exhibit similar properties. Moreover, the surrogate model is trained with latent-property pairs that are representative of the learned latent space. **(c)** During inversion, TrustMol optimizes a latent design by minimizing the predicted surrogate error and the epistemic uncertainty. The optimal latent design will then be decoded back into SELFIES by the pretrained SGP-VAE decoder.

Table 4. Effects of small perturbations on stability and property values. We randomly add $\mathcal{N}(0, 0.1)$ noise to an atom coordinate or a latent's component, and randomly change an atom type or a SELFIES' alphabet. We show the NFP errors between the original and perturbed molecules' properties.

Perturbation	Stable	NFP Error		
On	(%)	H (eV)	L (eV)	D (D)
Graph - 3D coord.	38.5	1.59	1.79	0.53
Graph - atom type	38.0	1.48	1.44	0.41
SELFIES	60.0	0.86	1.16	0.47
Latent	**67.2**	**0.42**	**0.47**	**0.24**

of stable molecules drastically decreases to 38.5%. Additionally, the properties of the remaining stable molecules changes significantly, as indicated by the relatively high MAE values. The same trend can be seen when the perturbation targets atom types of the 3D graphs, in which we randomly change a single atom type into another. Interestingly, utilizing SELFIES as molecular representations can improve robustness to such perturbations. For instance, replacing a randomly-selected alphabet in a SELFIES string with another valid alphabet only reduces the stability to 60.0%, while the MAEs between the original and perturbed molecular designs show improvements. Note that while SELFIES strings can always be translated into a stable molecule, the NFP that is used to generate the corresponding 3D conformation may not always converge due to the complexity of the molecule, which flags the molecule as unstable in our evaluation.

Finally, we can see that latent representations of molecules exhibit the greatest robustness toward perturbations. When a $\mathcal{N}(0, 0.1)$ noise is injected into the latents, the proportion of stable molecules remains high at 67.2%, and the MAE between the properties of the original and perturbed molecules is approximately 45% lower in average than that observed with SELFIES strings. These results validate our explanations regarding the high-frequency and discontinuous nature of the molecule-property mapping, and support our strategy of developing a custom latent space to smooth this mapping.