Supporting Information

PropMolFlow: Property-Guided Molecule Generation with Geometry-Complete Flow Matching

Cheng Zeng[1,2,†], Jirui Jin [1,2,†], Connor Ambrose[1,2], George Karypis[3], Mark Transtrum[4], Ellad B. Tadmor[5], Richard G. Hennig[2,6], Adrian Roitberg[1,2], Stefano Martiniani[7,8,9,10*], and Mingjie Liu[1,2*]

[1]*Department of Chemistry, University of Florida, Gainesville, FL 32611, USA*
[2]*Quantum Theory Project, University of Florida, Gainesville, FL 32611, USA*
[3]*Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN 55455, USA*
[4]*Department of Physics & Astronomy, Brigham Young University, Provo, UT 84602, USA*
[5]*Department of Aerospace Engineering and Mechanics, University of Minnesota, Minneapolis, MN 55455, USA*
[6]*Department of Materials Science & Engineering, University of Florida, Gainesville, FL 32611, USA*
[7]*Center for Soft Matter Research, Department of Physics, New York University, New York 10003, USA*
[8]*Simons Center for Computational Physical Chemistry, Department of Chemistry, New York University, New York 10003, USA*
[9]*Courant Institute of Mathematical Sciences, New York University, New York 10003, USA*
[10]*Center for Neural Science, New York University, New York 10003, USA*

December 13, 2025

[†]These authors contributed equally to this work
[*]sm7683@nyu.edu, mingjieliu@ufl.edu

Table of contents

Supplementary Table 1: Benchmark on the fixed SDF data using revised metrics. Results of PropMolFlow models trained on the previous original QM9 SDF data and the corrected data 'rQM9_v0.sdf'. A comparison for the commonly used stability metrics and the revised metrics is also provided. Mean and standard deviations are reported across all model checkpoints across different property embedding methods, properties and epochs.

Model	Atom Stable	rev Atom Stable	Mols Stable	rev Mols Stable	Closed-shell Ratio
Original data	99.7 ± 0.4	98.5 ± 0.4	95.9 ± 3.1	78.7 ± 4.3	91.1 ± 2.7
Corrected data	99.9 ± 0.0	99.7 ± 0.1	97.9 ± 0.7	95.2 ± 1.3	95.2 ± 1.3

1 QM9 SDF data issue

1.1 Results using corrected data and revised metrics

Supplementary Table 1 shows the effect of data correction and using the revised stability metrics. It is clear that the previous metrics substantially inflate the actual molecule stability (95.9% versus 78.7%) for models trained on the original QM9 SDF data. Moreover, the data correction improves the revised molecule stability from 78.7% and 95.2%, and the closed-shell ratios from 91.1% to 95.2%. The mean and standard deviations in this table were reported over all embedding methods, property types and model checkpoints.

1.2 Data Correction Procedure

The data-correction procedure described below corresponds to the latest rQM9_v1 SDF file available on the Zenodo repository [1]. The results presented in this work were generated using an earlier revision, rQM9_v0.sdf, which contained a larger number of unresolved molecules (935 versus 303 in rQM9_v1). As these additional corrections affect only a small fraction of the dataset, no performance differences are expected.

The original QM9 data in xyz format provides each molecule's SMILES and InChI strings (before quantum-mechanical relaxation) alongside its relaxed Cartesian coordinates. DeepChem's corresponding QM9 SDF file uses these relaxed coordinates, and, in addition, specifies explicit bond orders and atomic formal charges. Although all QM9 molecules are expected to be charge-neutral, numerous entries in the SDF exhibit nonzero net charges; even among those with zero net charge, many contain incorrect bond orders. Furthermore, a subset of molecules fails RDKit sanitization. Structural relaxation can also alter bond connectivity or stereochemistry, resulting in changes to the SMILES or InChI representations. Such geometry-induced SMILES changes were noted in the original QM9 publication [2] by comparing the input SMILES to those generated through OpenBabel program for relaxed geometries [3]. In total, 3054 molecules display this SMILES change, and previous generative models often exclude them entirely. Supplementary Table 2 lists the counts for each type of issue aforementioned.

Supplementary Table 2: Count of problematic molecules in the original QM9 SDF data. Except for the first two columns, there may be overlapping molecules in each category.

Non-zero net charges	Invalid bond orders with zero net charges	RDKit not sanitizable	Unmatched SMILES/InChI
25173	5174	1915	3054

Of the 3054 molecules whose SMILES no longer match after relaxation, 226 molecules form multiple fragments and must be discarded. However, a SMILES change alone does not imply an invalid structure. In our data-fix procedure, we retain any molecule whose relaxed geometry remains a single connected structure, and whose bond orders and formal charges can be assigned so that it is charge-neutral, passes RDKit sanitization, and satisfies valency-charge consistency, even if its SMILES has been altered by relaxation.

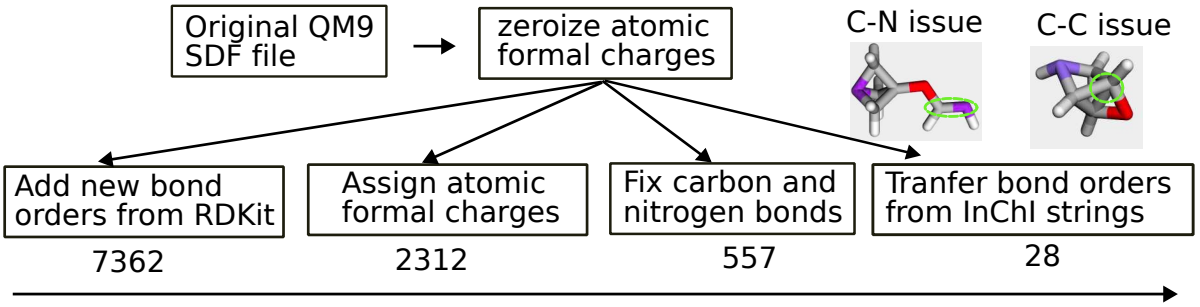
We used a cross-check workflow (Supplementary Figure 1) to assign bond orders and charges to each molecule. Starting from the DeepChem SDF, we take the following sequential steps; that is, the next step only addresses the remaining molecules from the previous step:

1. We zero all formal charges and rebuild bonds. Since non-zero charges were the most frequent issue, we first reset every atomic formal charge to zero and re-generated the bond connectivity with RDKit.

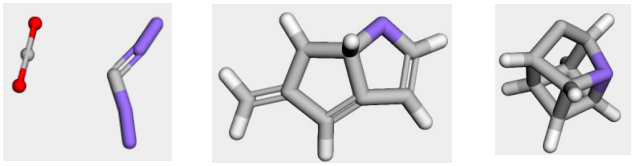
2. We keep the bond order from the original SDF. We adjust formal charges until each atom's valence matches its formal charge.

3. There are many remaining cases where C-N single bonds should be C=N or extraneous C-C bonds violate C(4) valency. Two such examples are provided in Supplementary Figure 1. These can be fixed by bond replacement and/or removal.

4. Finally, we transferred bond orders using the bond connectivity in the input InChI strings. This left just 28 unresolved molecules, which either fragmented after relaxation or for which no charge-neutral, valency-consistent assignment was possible (see three representative failures in Supplementary Figure 2).



Supplementary Figure 1: Flow chart to fix the bond and charge issues in the original QM9 SDF file. Two example configurations with C-N and C-C bond issues are provided along with the flowchart. Numbers below each step are the remaining molecules that are not fixed. C, N, O, and H are in gray, blue, red and white colors, respectively.



Supplementary Figure 2: Three examples of unresolved molecules. C, N, O, and H are in gray, blue, red and white colors, respectively. The three example molecules are among the 28 remaining molecules that are problematic and cannot be fixed with the data correction pipeline used in this work.

The bond and charge assignment procedure is given in the flowchart (Supplementary Figure 1. Since the major issue from Supplementary Table 2 is the non-zero charges, in the first step, we zeroize the atomic formal charges and add bonds using RDKit. For the remaining ones, we simply use the same bond orders in the original SDF file and only adjust the atomic formal charges to obtain structures with valid valency–charge configurations. Next, there are many molecules, where C-N single bonds should be replaced with double bonds and/or C-C single bonds should be removed to have the correct valency for N (3) and C (4). Two examples of the C-N and C-C bond issues are provided in Supplementary Figure 1, and the bond with issue is marked with a green dashed ellipse. Lastly, with bond transfer using input InChI strings, only 28 molecules remain. To ensure the success of bond transfers, we found that it is important to turn on the 'clearAromaticFlags' for 'Kekulization' of the RDKit molecule generated from the InChI string. Three examples of the unresolved cases are provided in Supplementary Figure 2, where either molecules break apart with relaxation or there is no way to ensure charge neutrality and valency validity with the relaxed structures.

Next, we compare the post-adjustment InChI strings (with stereochemistry removed) against the original InChIs from the XYZ files, since omitting stereochemical layers preserves bond orders and formal charges. Excluding stereochemistry, only 945 molecules remain unmatched. Of these, 696 molecules lie among the original 3054 "should-not-match" cases, reflecting connectivity changes on relaxation.

The remaining 249 molecules fall outside that set; 51 of them fail RDKit sanitization and are discarded. For the other 198, both the original and adjusted InChIs are valid. We compute each variant's MMFF94 potential energy in RDKit to decide which bonding is more plausible. Eighty adjusted-bond molecules exhibit energies no greater than their originals, so we keep their new bond/charge assignments; three molecules cannot be parameterized by MMFF94 and we used the assignments. The final 116 molecules adopt the bond orders transferred from the original InChI strings, and this completes our bond- and charge-correction pipeline for the QM9 SDF, which is available at [1].

The differences between 'rQM9_v0' and 'rQM9_v1' SDF files come from the step 3 and 4. For the earlier version, we did not consider molecules which have both N- and C-bond issues in step 3, and we did not clear aromatic flags in the molecules in step 4. This would lead to a smaller, but still valid dataset for training PropMolFlow models and GVP predictors after data precessing.

2 Systematic comparison of property embedding methods

2.1 Summary of top-performing methods

For α and C_v, the lowest MAEs in both OOD and ID tasks are achieved with the 'Sum' operation without Gaussian expansion. In contrast, for μ, the best performance is obtained using 'Multiply' and 'Sum' for respective OOD and ID tasks. For molecular orbital-related properties such as $\Delta\epsilon$, ϵ_{HOMO} and ϵ_{LUMO}, the best performance is typically achieved with a concatenation-based operation.

Supplementary Table 3: Optimal property embedding methods by task and property. The best methods are identified by comparing the MAEs in corresponding tasks using a GVP property predictor.

	ID Tasks		OOD Tasks	
Property	Embedding method	Gaussian expansion	Embedding method	Gaussian expansion
α	Sum	No	Sum	No
$\Delta\epsilon$	Concatenate	No	Concatenate_Sum	No
ϵ_{HOMO}	Concatenate_Sum	No	Concatenate	No
ϵ_{LUMO}	Concatenate_Multiply	No	Concatenate_Sum	No
μ	Sum	No	Multiply	No
C_v	Sum	No	Sum	No

To compare the overall effect of different embedding methods, ranges of MAEs across all saved checkpoint models are listed in Supplementary Table 4.

Supplementary Table 4: Lowest and highest MAEs for each property by varying the property embedding methods and saved model checkpoints in the ID tasks. The MAEs are calculated by comparing the GVP-predicted values to the input target values.

Property	α	$\Delta\epsilon$	ϵ_{HOMO}	ϵ_{LUMO}	μ	C_v
Unit	Bohr^3	meV	meV	meV	Debye	$\text{cal/(mol} \cdot \text{K)}$
Minimum MAE	1.31	391	254	315	0.620	0.626
Maximum MAE	1.77	551	324	437	0.836	0.847

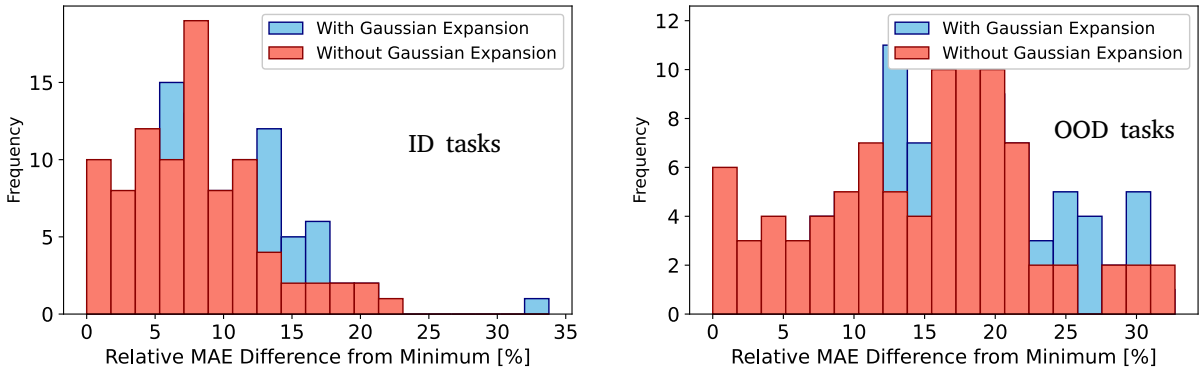
2.2 Relative MAEs

To further assess the impact of different property embedding strategies and the use of Gaussian expansion, we analyze statistics of relative MAE differences for all models. This relative MAE difference is the difference between the MAE for a specific model on a given property and the minimum MAE for that property. Comparing the histograms in Supplementary Figure 3, it is confirmed from these results that Gaussian expansion should not be included for top-performing models in the OOD and ID tasks, even though given the same property type and property-embedding operation, Gaussian expansion leads to a lower MAE for 8 out of 30 cases for ID tasks and 12 out of 30 cases for OOD tasks.

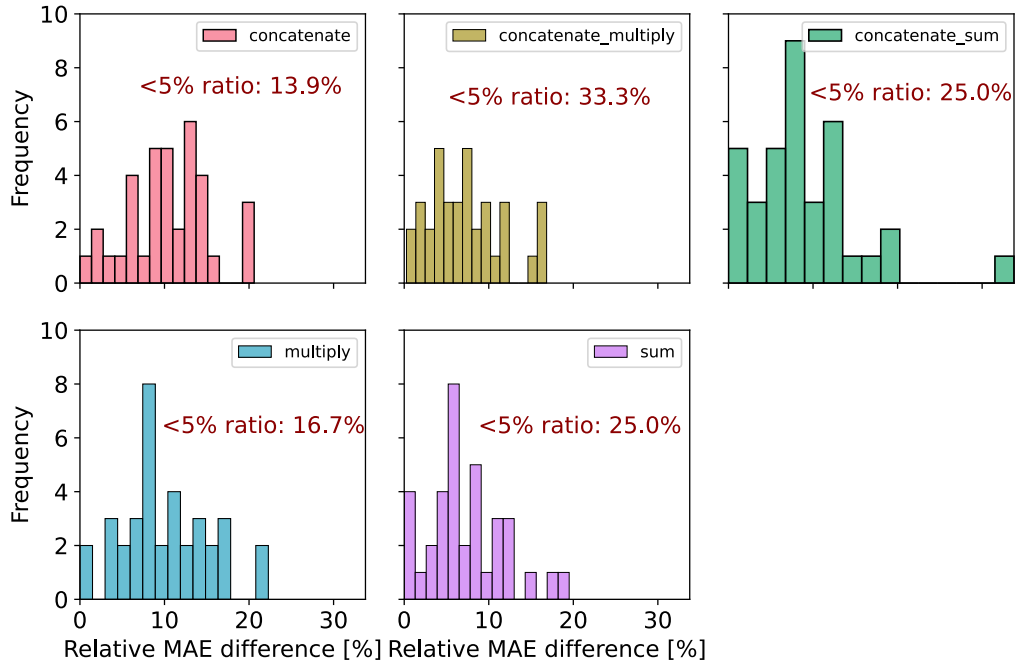
This difference may be attributed to how input properties interact with the node scalar features. Without Gaussian expansion, the model freely explores the latent space, and the learned correlations between structure and molecular properties are more influenced by the natural distribution

of the QM9 data. While Gaussian expansion is demonstrated to be effective for autoregressive models [4], it uses rigid functional forms that may distort the structure and property relationship towards undesirable directions, which leads to a slightly inferior performance compared to results without Gaussian expansion.

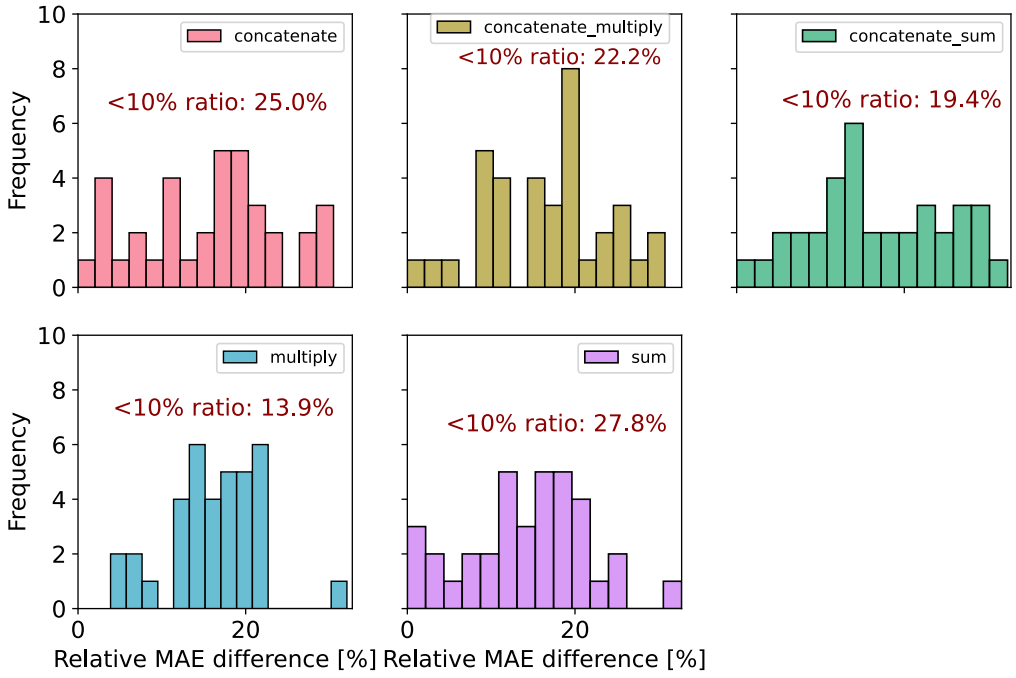
Regarding the property embedding methods themselves, results are summarized in separate histograms in Supplementary Figure 4 and in Figure 5. With a cutoff of 10% and 5% for respective OOD and ID tasks, the "Sum" operation appears more frequently, followed by Concatenate and Concatenate_Multiply for the OOD tasks, while Concatenate_Multiply and Sum operations perform better in the ID tasks. Based on these findings, we recommend that, in the absence of prior knowledge, Sum or Concatenate_Sum property embedding without Gaussian expansion is likely to yield favorable results for property-guided generation.



Supplementary Figure 3: Histogram of relative MAE differences from the minimum for each property in the ID (left) and OOD (right) tasks. Results are for two groups: One with Gaussian expansion and the other without Gaussian expansion.



Supplementary Figure 4: Histogram of relative MAE differences from the minimum for each property based on different property embedding methods in the ID tasks. The relative MAE difference is the difference between the MAE for an embedding method on a given property and the minimum MAE for that property, divided by the minimum MAE values. In this way, we can group the relative MAEs across different property types. For each embedding method, results with the lowest 3 MAEs are included in this comparison.



Supplementary Figure 5: Histogram of relative MAE differences from the minimum for each property based on different property embedding methods in the OOD tasks. The relative MAE difference is the difference between the MAE for an embedding method on a given property and the minimum MAE for that property, divided by the minimum MAE values. In this way, we can group the relative MAEs across different property types. For each embedding method, results with the lowest 3 MAEs are included in this comparison.

3 Optimal transport alignment

Previous works have shown that aligning the prior and target samples via equivariant optimal transport can improve the performance of flow matching by minimizing the cross-links of the conditional probability paths [5]. For single molecule generation, optimal transport alignment can be realized through computing the optimal permutation of node ordering and rigid body alignment of atomic positions [6, 7]. FlowMol applies the alignment to atomic positions for the training process. Additionally, this alignment ensures that the atomic positions exist in the center-of-mass free space, as proposed by Hoogeboom et al. [8], which guarantees to give translation invariance.

4 Proofs of probability invariance for an equivariant generative process

To be self-contained, we show in the following that an invariant prior probability distribution p_0, along with an equivariant transition probability path $p(x_{t+\Delta t}|x_t)$, leads to an invariant marginal probability $p(x_{t+\Delta t})$. Klein et al. further proved that optimal transport preserves the symmetry of both the probability state and the transport path [7].

4.1 O(3) invariance of the prior distribution.

We can observe that $p(x_0) = \mathcal{N}(0, \mathbf{I}_3)$ to be invariant for rotations and reflections, so $p(x_0) = p(Rx_0)$ where $R \in \mathbb{R}^{3\times3} : RR^T = I$.

4.2 O(3) invariance of the marginal probability.

Suppose that we now have an equivariant transition probability path, which leads to $p(Rx_t|Rx_p) = p(x_t|x_p)$. If $p(x_t)$ is invariant and the transition probability path is equivariant for rotations and reflections, using the proof by Xu et al. [9] and Hoogeboom et al. [8], we have:

$$
\begin{aligned}
p(Rx_{t+\Delta t}) &= \int_{x_t} p\left(Rx_{t+\Delta t}|x_t\right) p(x_t) dx_t && \text{Probability chain rule} \\
&= \int_{x_t} p\left(Rx_{t+\Delta t}|RR^{-1}x_t\right) p(RR^{-1}x_t) dx_t && \text{Multiply by } RR^{-1} = I \\
&= \int_{x_t} p\left(x_{t+\Delta t}|R^{-1}x_t\right) p(R^{-1}x_t) dx_t && \text{Equivariance \& Invariance} \\
&= \int_{x_t} p\left(x_{t+\Delta t}|u\right) p(u) |\underbrace{\det\left(R\right)}_{=1}| du && \text{Change of variable } u = R^{-1}x_t \\
&= p(x_{t+\Delta t}) && \text{Definition of marginal probability}
\end{aligned}
$$

Thus, the marginal probability $p(x_{t+\Delta t})$ is also invariant to rotations.

5 SE(3) equivariant neural networks

As we know, chiral molecules are mirrored along a stereocenter, axis, or plane. Since molecule properties can heavily rely on the chirality, SE(3) GVP neural networks, which are sensitive to reflection operations, were used. This is achieved by introducing cross-product operations of vectors in the GVP updates (Algorithm 1 of Ref. [10]). The resulting SE(3) graph neural networks are hence sensitive to reflections while preserving the rotational equivariance. By design, the atomic coordinate updates are a combination of operations on relative position vectors and cross-products of positive vectors. Using a proof similar to Schneuing et al. [11], we show below that an update involving relative position vector is E(3) equivariant while an update involving vector cross products is SE(3) equivariant. A combination of both operations is hence an SE(3)-equivariant process.

Suppose we have a rotation or reflection operation defined by an orthogonal matrix $R \in \mathbb{R}^{3\times3}$ with $R^TR = \mathbf{I}$, and a translation vector $t \in \mathbb{R}^3$ Given the transformation of $\widehat{x}_i = Rx_i + t$ and $\widehat{x}_j = Rx_j + t$, the new relative position vector changes to:

$$
\widehat{x}_i - \widehat{x}_j = R(x_i - x_j) \tag{1}
$$

Hence, operations involving the relative position vector are equivariant to rotation/reflection.

For a cross product of relative position vectors, we let the original vector product be $(x_i - \overline{x}) \times (x_j - \overline{x})$. The transformation using R and t changes the cross-product to:

$$
\begin{aligned}
(\widehat{x}_i - \overline{\widehat{x}}) \times (\widehat{x}_j - \overline{\widehat{x}}) &= (Rx_i - R\overline{x}) \times (Rx_j - R\overline{x}) && \text{Cancellation of translation} \\
&= (R(x_i - \overline{x})) \times (R(x_j - \overline{x})) && \text{Factoring out R} \\
&= \det(R) R((x_i - \overline{x}) \times (x_j - \overline{x}))) && (Ra) \times (Rb) = \det(R) R(a \times b)
\end{aligned}
$$

As a result, the cross-product vector is equivariant to rotation where $\det(R) = 1$ while the symmetry breaks for reflection operations where $\det(R) = -1$.

6 Supplementary results

6.1 Numeric results for six structural validity metrics

All the numbers correspond to the results shown in Fig. 2.

Supplementary Table 5: Molecular generation results on QM9. All metrics use the percentage (%) as the unit, and higher numbers indicate better performance. Mean and standard deviations of all metrics are reported across six properties for sampling 10000 structures. PropMolFlow results are averaged across all properties for the top-performing models shown in Supplementary Table 3. All results are based on our sampling. Note that the PropMolFlow numbers are slightly different from those in Supplementary Table 1 because this one uses the top-forming models and the previous table uses all the models. Best results are in bold.

Model	Atom Stable	Mols Stable	Mols Valid	Valid & Unique	PB-valid	Closed-shell Ratio
EEGSDE	98.1 ± 0.1	79.9 ± 0.7	90.3 ± 0.8	89.5 ± 0.8	87.1 ± 0.4	86.5 ± 0.8
GCDM	98.7 ± 0.1	85.7 ± 1.4	94.6 ± 0.5	93.6 ± 0.5	92.0 ± 0.5	90.7 ± 1.3
GeoLDM	98.4 ± 0.1	83.1 ± 1.4	91.7 ± 1.1	90.9 ± 1.0	89.4 ± 1.0	89.4 ± 1.0
EquiFM	93.4 ± 0.6	40.4 ± 4.0	78.3 ± 2.5	78.1 ± 2.5	64.8 ± 3.2	59.9 ± 2.0
JODO	99.2 ± 0.1	93.0 ± 0.8	96.2 ± 0.5	94.3 ± 0.5	94.9 ± 0.5	$\mathbf{98.4 \pm 0.5}$
PropMolFlow	$\mathbf{99.7 \pm 0.1}$	$\mathbf{95.2 \pm 1.0}$	$\mathbf{97.8 \pm 0.8}$	$\mathbf{95.5 \pm 1.1}$	$\mathbf{96.5 \pm 1.1}$	94.8 ± 1.2

When evaluating metrics on a per-property basis, PropMolFlow consistently achieves the highest molecular stability and validity across all properties (Supplementary Tables 6-10). Across properties, PropMolFlow's metrics follow the order: $\mu \approx \Delta\epsilon > \epsilon_{\text{HOMO}} > \epsilon_{\text{LUMO}} > \alpha > C_v$. This trend generally holds for the baseline models as well, with the exception of EEGSDE, suggesting a more complex structure–property relationship for α and C_v. This complexity may be due to their strong dependence on molecular geometry and the limitation of the training data, which includes only fully relaxed molecules. Across metrics, molecular stability and "Validity & Uniqueness" exhibit over 2% variations across properties, whereas molecular validity and PoseBusters validity vary by less than 1%. Besides, all methods in this study generate a non-negligible fraction of open-shell molecules: on average 1.6% for JODO and at least 4% for the other methods. This highlights a previously overlooked issue in generative molecular modeling: the inability to consistently respect the valence electron constraints inherent to the training data. Addressing this limitation is essential

for improving the chemical validity of generated molecules especially when atomic formal charges are considered in molecular graphs, and future work should improve upon this direction—perhaps by explicitly including closed-shell constraints, as we do here by adding bond orders into molecular graphs—to ensure better alignment with the reference distribution.

6.1.1 Per-property structural validity results for PropMolFlow

Supplementary Table 6: Per-property molecular generation results on QM9 for PropMolFlow. Results are reported using the top-performing model for each property and evaluated using the revised stability metrics.

PropMolFlow	Atom Stable	Mols Stable	Mols Valid	Valid & Unique	PB-valid	Closed-shell Ratio
α	99.6	93.6	96.5	94.7	92.9	94.7
$\Delta\epsilon$	99.7	94.9	97.8	95.9	94.8	96.4
ϵ_{HOMO}	99.8	96.6	97.7	94.2	96.3	96.8
ϵ_{LUMO}	99.8	96.0	98.9	96.8	95.7	97.8
μ	99.7	95.4	97.5	95.0	95.1	96.1
C_v	99.7	95.0	98.4	96.7	93.8	97.3

6.1.2 Per-property structural validity results for EEGSDE

Supplementary Table 7: Per-property molecular generation results on QM9 for EEGSDE models. Results are based on our own sampled molecules using publicly available generative models and evaluated using the revised stability metrics.

EEGSDE	Atom Stable	Mols Stable	Mols Valid	Valid & Unique	PB-valid	Closed-shell Ratio
α	98.1	79.6	88.5	88.0	86.3	86.7
$\Delta\epsilon$	98.1	79.8	90.5	89.6	87.2	87.0
ϵ_{HOMO}	98.1	78.7	91.2	90.4	87.2	84.7
ϵ_{LUMO}	98.2	80.9	90.7	90.0	87.7	87.2
μ	98.1	79.7	90.5	89.7	86.8	86.4
C_v	98.2	80.5	90.1	89.3	87.2	87.1

6.1.3 Per-property structural validity results for GCDM

Supplementary Table 8: Per-property molecular generation results on QM9 for GCDM models. Results are based on our own sampled molecules using publicly available generative models and evaluated using the revised stability metrics.

GCDM	Atom Stable	Mols Stable	Mols Valid	Valid & Unique	PB-valid	Closed-shell Ratio
α	98.6	85.0	94.4	93.7	91.6	89.9
$\Delta\epsilon$	98.7	86.0	94.7	93.9	92.0	90.9
ϵ_{HOMO}	98.9	88.3	95.4	94.3	92.9	93.1
ϵ_{LUMO}	98.7	84.7	94.4	93.5	92.0	89.4
μ	98.7	86.3	94.9	93.7	92.4	91.5
C_v	98.7	85.1	94.4	93.7	91.4	89.9

6.1.4 Per-property structural validity results for GeoLDM

Supplementary Table 9: Per-property molecular generation results on QM9 for GeoLDM models. Results are based on our own sampled molecules using the checkpoint models bundled with the GCDM work [12] and evaluated using the revised stability metrics.

GeoLDM	Atom Stable	Mols Stable	Mols Valid	Valid & Unique	PB-valid	Closed-shell Ratio
α	98.2	81.4	91.6	90.8	89.1	87.9
$\Delta\epsilon$	98.3	83.1	91.8	90.9	89.2	89.9
ϵ_{HOMO}	98.4	84.0	92.2	91.5	90.3	90.4
ϵ_{LUMO}	98.4	84.0	92.2	91.4	89.9	89.3
μ	98.6	85.5	93.0	92.2	90.3	90.8
C_v	98.2	81.3	89.6	88.8	87.3	88.3

6.1.5 Per-property structural validity results for JODO

Supplementary Table 10: Per-property molecular generation results on QM9 for JODO models. Results are based on our own sampled molecules using publicly available generative models and evaluated using the revised stability metrics.

JODO	Atom Stable	Mols Stable	Mols Valid	Valid & Unique	PB-valid	Closed-shell Ratio
α	99.1	92.7	96.4	94.4	95.2	98.7
$\Delta\epsilon$	99.3	94.1	97.0	95.1	95.7	99.0
ϵ_{HOMO}	99.2	93.5	95.9	94.0	94.7	98.5
ϵ_{LUMO}	99.1	92.5	95.6	93.9	94.5	98.2
μ	99.3	93.7	96.4	94.7	95.3	98.7
C_v	99.1	91.7	95.6	93.9	94.1	97.4

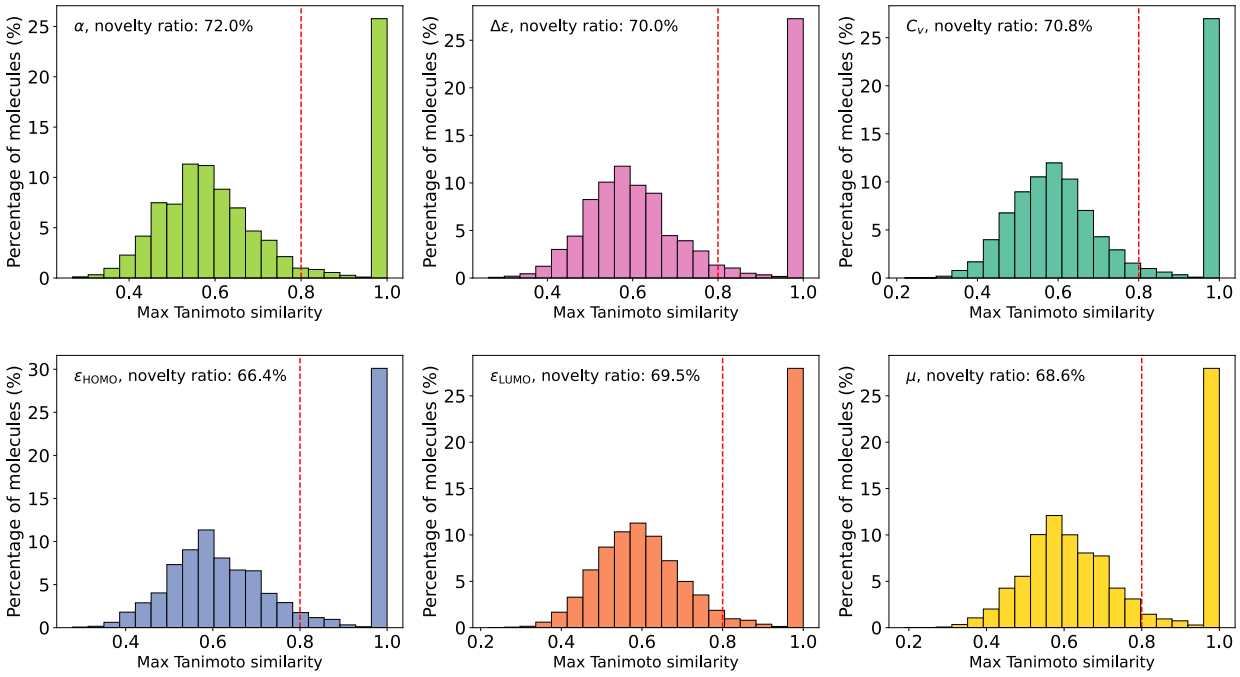
6.1.6 Per-property structural validity results for EquiFM

Supplementary Table 11: Per-property molecular generation results on QM9 for EquiFM models. Results are based on our own sampled molecules using models trained by us based on the hyperparameters in Supplementary Section 8.2 and evaluated using the revised stability metrics.

EquiFM	Atom Stable	Mols Stable	Mols Valid	Valid & Unique	PB-valid	Closed-shell Ratio
α	93.2	39.2	76.9	76.8	65.4	58.1
$\Delta\epsilon$	93.5	41.7	80.2	80.1	65.9	60.0
ϵ_{HOMO}	93.8	42.7	80.8	80.6	68.1	62.4
ϵ_{LUMO}	93.6	41.3	77.6	77.5	64.9	60.2
μ	92.2	32.4	73.7	73.6	58.0	56.5
C_v	94.1	45.0	80.4	80.2	66.6	61.9

6.2 Maximum Tanimoto similarity of generated molecules compared to training data in the ID tasks

Supplementary Figure 6 shows the maximum Tanimoto similarity of PropMolFlow-generated molecules in the ID task with respect to the training set. Similarities were computed using the Jaccard index on Morgan fingerprints, as implemented in RDKit [13].

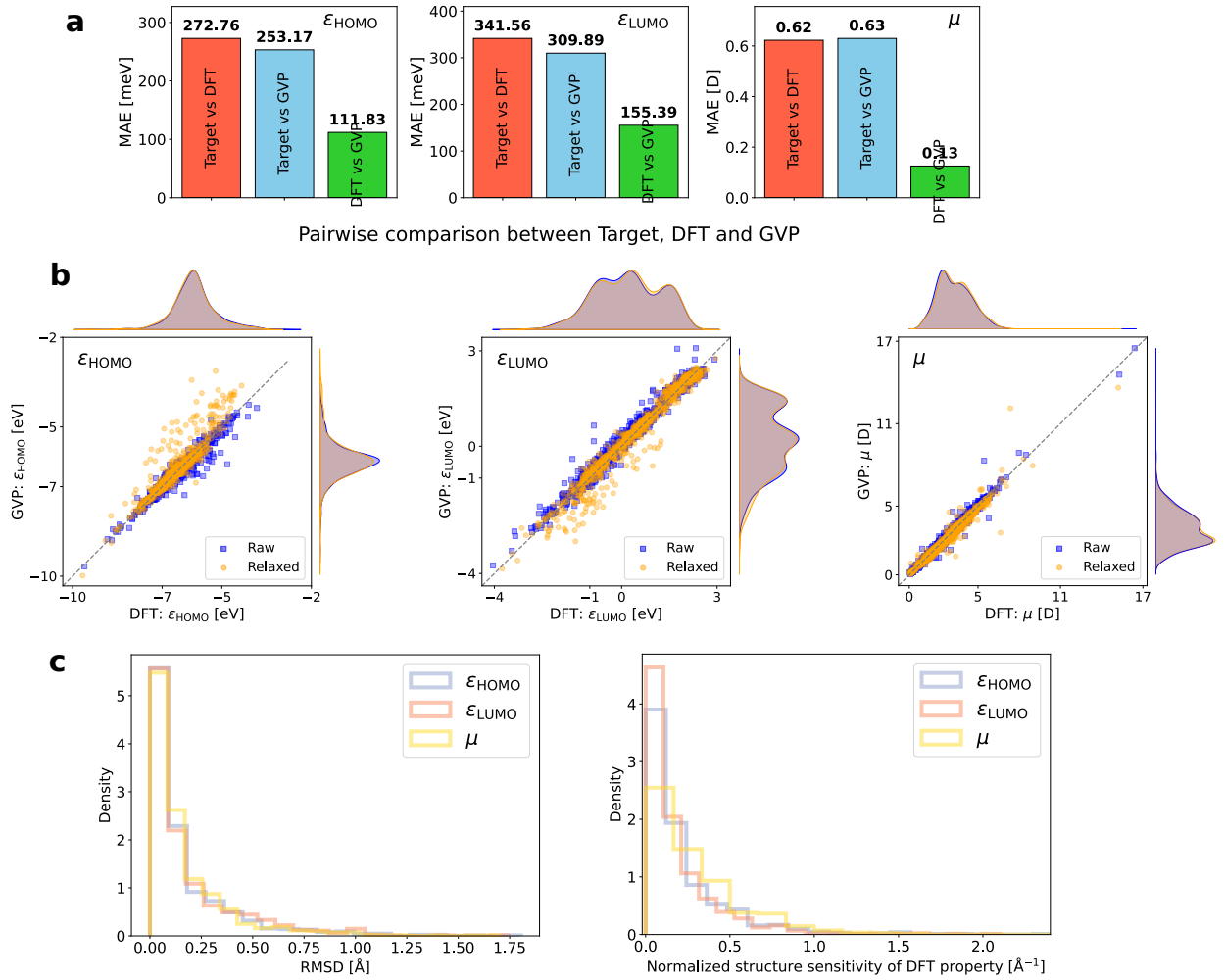


Supplementary Figure 6: Pecentages of molecules with different maximum Tanimoto similarities among generated and filtered molecules compared to training data in the ID tasks. Maximum Tanimoto similarity of generated molecules after RDKit sanitization are evaluated using a Morgan fingerprint with 2048 Bits [13]. Dashed lines indicate the cutoff of 0.8 to define novel rations.

6.3 Evaluation of GVP property predictors

6.3.1 Performance of GVP property predictors for ϵ_{HOMO}, ϵ_{LUMO} and μ

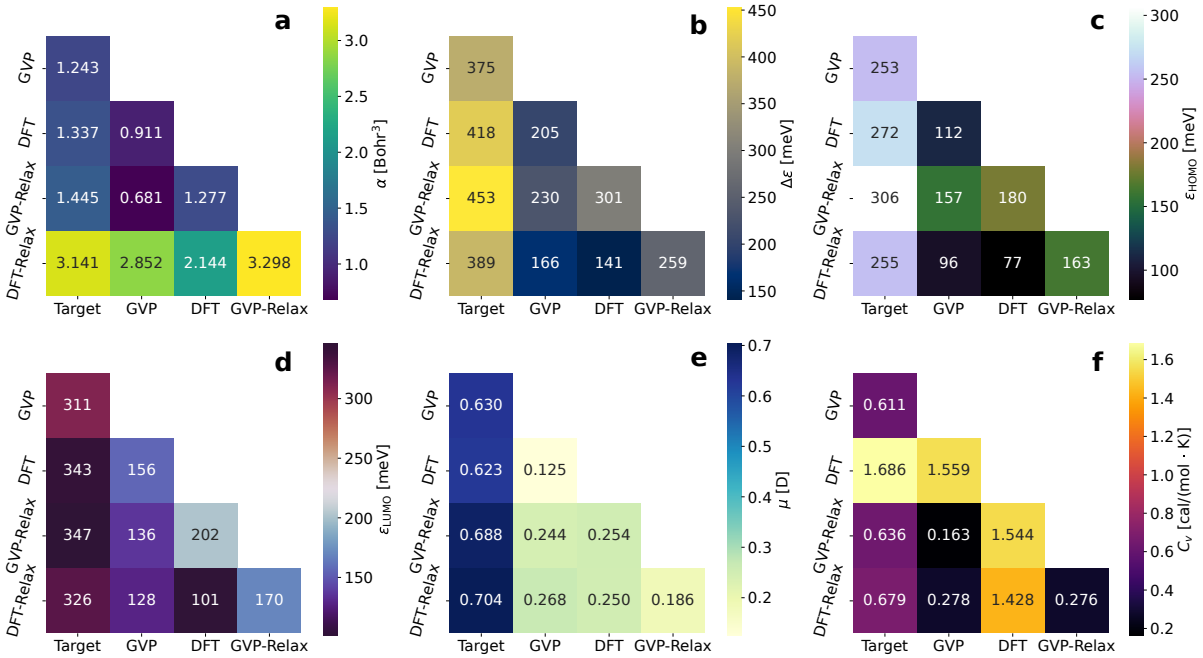
Detailed comparison for ϵ_{HOMO}, ϵ_{LUMO} and μ are depicted in Supplementary Figure 7. For ϵ_{HOMO} and ϵ_{LUMO}, the results are very similar to that of $\Delta\epsilon$. Results for μ are slightly different in that its GVP versus Target MAE matches well with that of DFT. Also, μ shows a slightly higher structural dependence of DFT property values than the other two orbital-base properties.



Supplementary Figure 7: Performance of GVP property predictors without and with relaxation for μ, ϵ_{HOMO} and ϵ_{LUMO}. **a**, Pairwise comparison between Target, DFT and GVP on raw molecules. **b**, GVP versus DFT for both raw and DFT-relaxed molecules. **c**, Structure variation and property dependence on the DFT relaxation.

6.3.2 Pairwise MAEs between DFT, GVP and Target

Pairwise MAEs between Target, DFT and GVP are shown in Supplementary Figure 8 for six properties without and with relaxation. Results evaluated on DFT-relaxed molecules are indicated by '-Relax'.



Supplementary Figure 8: Pairwise MAEs between DFT, GVP and Target with and without relaxation. Each heatmap corresponds to one specific molecule property: **a**, Isotropic polarizability (α); **b**, HOMO-LUMO energy gap ($\Delta\epsilon$); **c**, HOMO energy (ϵ_{HOMO}); **d**, LUMO energy (ϵ_{LUMO}); **e**, Dipole moment (μ); **f**, Room-temperature heat capacity (C_v). 'target' represents the input property values used to generate molecules using PropMolFlow models. '-Relax' indicates that the evaluation is performed on the DFT relaxed structures.

6.4 Interpolation results from the PropMolFlow models

6.4.1 Numeric results for interpolation study

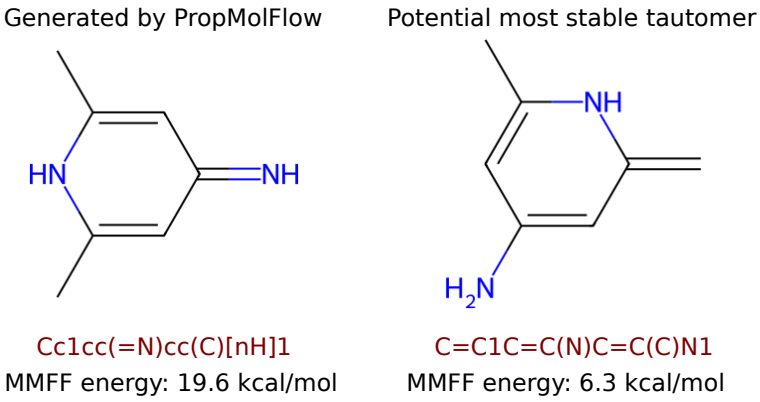
Supplementary Table 12 shows the input target property values, the DFT-computed values and the GVP predictions corresponding to the molecule configurations in Fig. 4.

Supplementary Table 12: Input target property values, DFT-computed values and GVP predictions for the interpolation study. α, $\Delta\epsilon$, ϵ_{HOMO}, ϵ_{LUMO}, μ, and C_v use Bohr^3, eV, eV, eV, Debye, and cal/(mol · K) as the respective units.

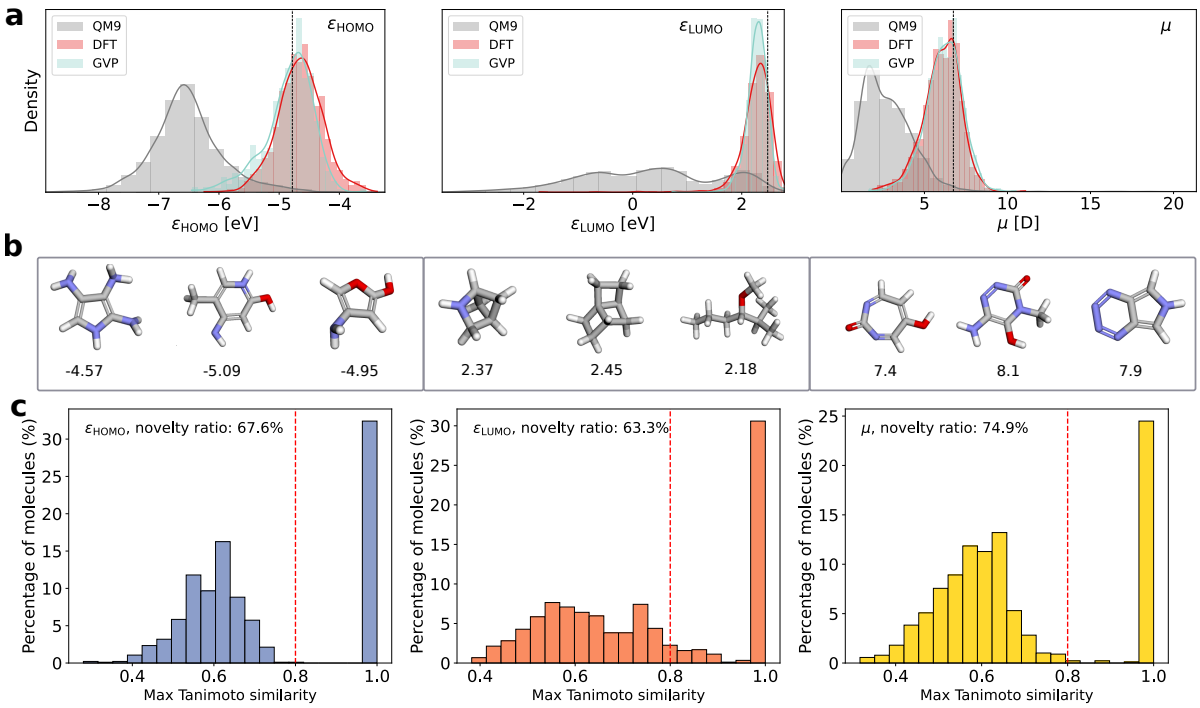
Type, Property	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8
Target, α	63.8	68.4	72.9	77.5	82.0	86.6	91.1	95.7
DFT, α	63.9	68.4	72.9	77.3	81.9	86.6	90.6	95.6
GVP, α	64.8	69.7	74.1	78.7	84.1	88.5	92.6	95.7
Target, $\Delta\epsilon$	5.03	5.77	6.47	7.17	7.92	8.63	9.33	10.1
DFT, $\Delta\epsilon$	5.14	5.72	6.53	7.17	8.02	8.66	9.30	10.0
GVP, $\Delta\epsilon$	5.25	5.90	6.53	7.24	8.11	8.90	9.09	10.1
Target, ϵ_{HOMO}	-7.37	-7.02	-6.69	-6.36	-6.04	-5.69	-5.36	-5.03
DFT, ϵ_{HOMO}	-7.35	-7.02	-6.69	-6.34	-6.12	-5.71	-5.55	-5.03
GVP, ϵ_{HOMO}	-7.35	-6.89	-6.66	-6.36	-6.23	-5.55	-5.60	-4.85
Target, ϵ_{LUMO}	-2.88	-2.18	-1.47	-0.76	-0.05	0.65	1.36	2.07
DFT, ϵ_{LUMO}	-2.86	-2.10	-1.39	-0.79	-0.19	0.65	1.39	2.10
GVP, ϵ_{LUMO}	-2.69	-2.18	-1.47	-0.87	-0.14	0.82	1.52	2.04
Target, μ	0.580	1.43	2.28	3.13	3.98	4.83	5.68	6.53
DFT, μ	0.675	1.37	2.19	3.14	3.95	4.85	5.40	6.69
GVP, μ	0.578	1.31	2.29	3.21	4.08	4.92	5.06	6.69
Target, C_v	23.7	26.0	28.2	30.5	32.7	35.0	37.2	39.5
DFT, C_v	24.1	26.1	28.1	30.5	32.7	35.3	37.4	36.9
GVP, C_v	23.9	27.5	28.3	30.2	33.1	35.7	37.4	38.3

6.4.2 Comparing two isomers

Supplementary Figure 9 shows the comparison between the PropMolFlow-generated molecule conditioned on ϵ_{HOMO} versus the most stable tautomer identified by RDKit and MMFF empirical energy calculator.



Supplementary Figure 9: PropMolFlow-generated molecule versus the most stable counterpart identified by RDKit with Merck Molecular Force Fields (MMFF) potential energies. The 2D molecular graph, SMILES and the MMFF-calculated potential energies are compared. RDKit 'TautomerEnumerator' was used to find potential tautomers.



Supplementary Figure 10: Performance of PropMolFlow in OOD generation for ϵ_{HOMO}, ϵ_{LUMO} and μ. a, Distribution of DFT calculated and GVP predicted values for PropMolFlow generated molecules, and the property distribution of QM9 is also included. The vertical black dashed line in histograms represents the target property value $q_{0.99}$. Curves on top of histograms are fitted with a kernel density estimation. **b**, Three example molecules that do not exist in QM9 but are found in a larger PubChem dataset are included in the left panel. Numbers below the configurations are DFT calculated property values on raw molecules generated by PropMolFlow models. C, H, O, N, and F are in gray, white, red, blue, and yellow colors, respectively. Property values for ϵ_{HOMO}, ϵ_{LUMO}, and μ are in units of eV, eV and Debye, respectively. **c**, Maximum Tanimoto similarity of generated molecules compared to the training data using a Morgan fingerprint with 2048 Bits [13].

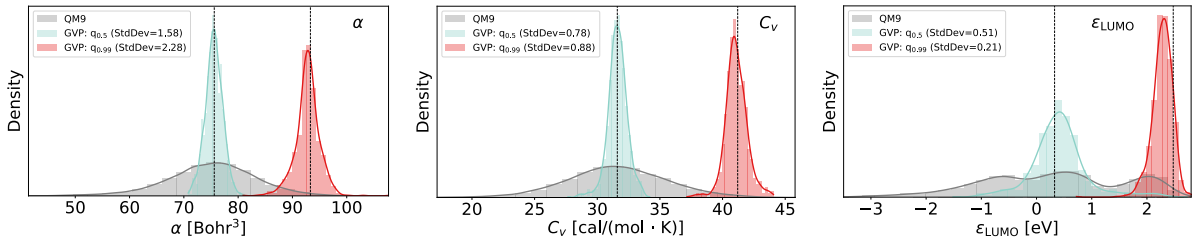
6.5 Toward OOD generation

6.5.1 Performance conditioned on ϵ_{HOMO}, ϵ_{LUMO} and μ

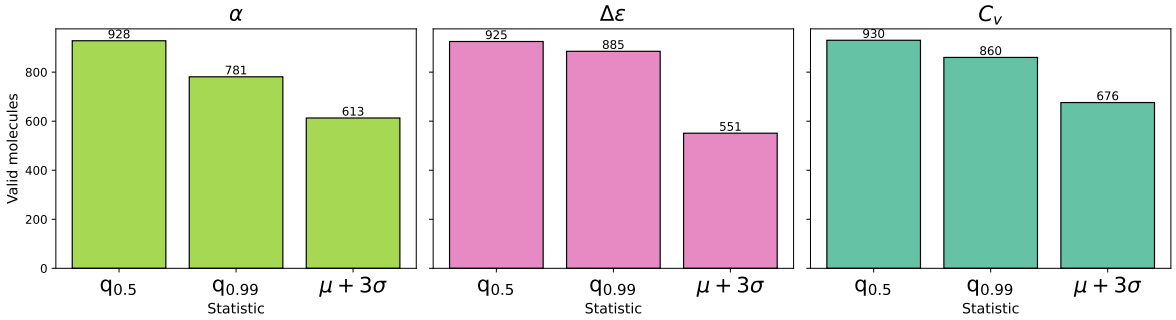
The OOD generation results for ϵ_{HOMO}, ϵ_{LUMO} and μ are shown in Supplementary Figure 10.

6.5.2 Additional extrapolation experiments

In Fig. 5, we observed an offset for the center of distribution against the target. As a comparison, we also sampled the same number of molecules using the median values of QM9 training data ($q_{0.5}$), which are closer to the center of the distribution. As shown in Supplementary Figure 11, the distribution of GVP-predicted property values for generated molecules are typically narrower than those obtained with a more extreme target property value ($q_{0.99}$), when the training data are approximately normal, such as for α, μ, C_v, and ϵ_{HOMO}. In contrast, for properties with broader training data distributions, such as $\Delta\epsilon$ and ϵ_{LUMO}, the $q_{0.5}$ case produces broader GVP distributions. This suggests that the quality of generation is strongly influenced by the underlying distribution of



Supplementary Figure 11: Comparison of GVP-predicted values for molecules conditioned on median values ($q_{0.5}$) and the 99th quantile ($q_{0.99}$) respectively. Results for three example properties are shown, including α, C_v and ϵ_{LUMO}.



Supplementary Figure 12: Numbers of valid molecules when conditioning on different property values, including $q_{0.5}$, $\mu + 3\sigma$ and $q_{0.99}$. Results for three example properties are shown, including α, C_v and $\Delta\epsilon$. Note that the μ in this figure corresponds to the mean value of QM9 training data, not the property dipole moment.

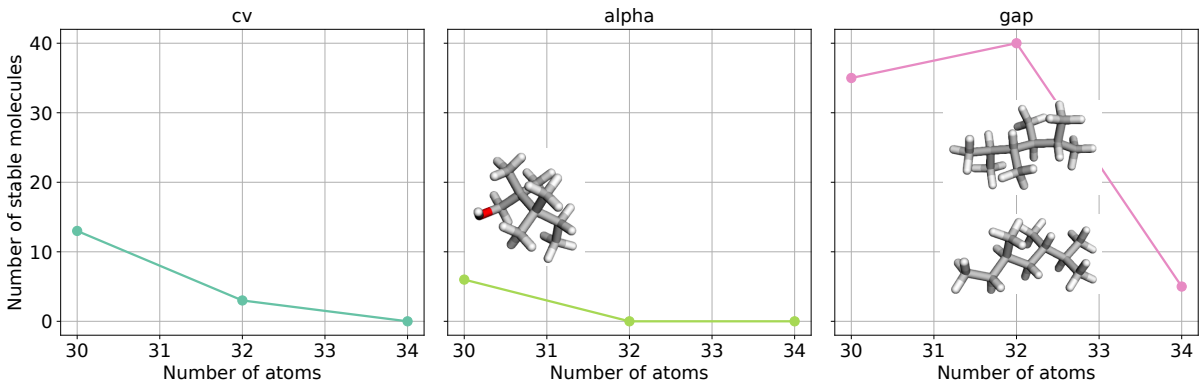
the training data.

To further evaluate the model's extrapolation capability, we examined molecule generation at extreme atom counts and property values. While the model can still produce valid molecules, the fraction of valid outputs decreases as property values (Supplementary Figure 12) and atom counts (Supplementary Figure 13) move further beyond the training regime, highlighting a fundamental limitation imposed by the available training data.

The specific property values for each bar in Supplementary Figure 12 are shown in Supplementary Table 13. Supplementary Figure 13 shows the total number of valid molecules when conditioning on ten different property values for ten sweeps (100 molecules in total), and note that the maximum atom count in QM9 is 29.

Supplementary Table 13: Three selected values for each property. The 'μ' here represents the mean of training data property values rather than the dipole moment elsewhere.

Statistic	$\Delta\epsilon_{\text{HOMO}}$ (eV)	$\Delta\epsilon_{\text{LUMO}}$ (eV)	$\Delta\epsilon$ (eV)	μ (D)	α (bohr^3)	C_v (cal·mol^{-1}·K^{-1})
$q_{0.5}$ (median)	-6.560	0.327	6.800	2.500	75.60	31.600
$q_{0.99}$ (99th pct.)	-4.781	2.476	9.350	6.742	93.31	41.208
$\mu + 3\ \sigma$	-4.626	4.082	10.612	7.300	100.00	44.000



Supplementary Figure 13: Numbers of valid molecules when conditioning on atom counts beyond the training data. A few examples configurations are provided in the inset. C, H, and O are in grey, white and red color, respectively. Lines are included to guide the eye.

7 Analytics for the curated DFT dataset

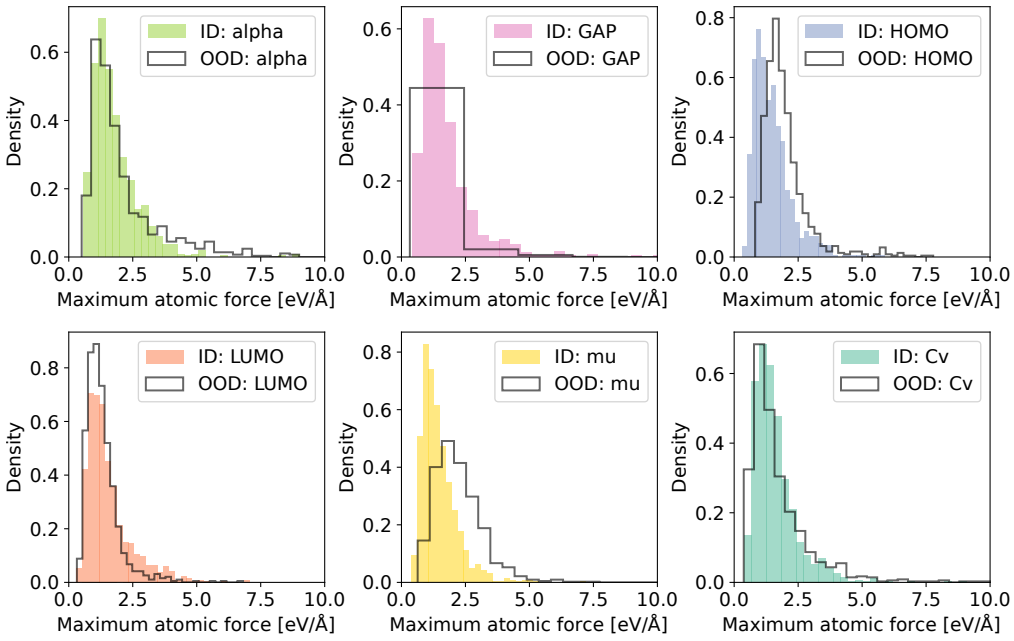
7.1 Number of configurations in the curated DFT data

Supplementary Table 14 shows that we normally have more remaining configurations after filtering for the ID tasks compared to the OOD tasks, except for $\Delta\epsilon$. We find HOMO, LUMO and dipole moment based PropMolFlow models generate more valid structures compared to polarizability and heat capacity, in agreement with structural validity results (Supplementary Table 6), except for $\Delta\epsilon$.

Supplementary Table 14: Number of remaining valid molecules after molecule filtering. Molecular filtering steps with the criteria, including Molecule stability, RDKit validity, PoseBusters validity, closed-shell validity, multi-fragment check, and DFT convergence.

Task Type	α	$\Delta\epsilon$	ϵ_{HOMO}	ϵ_{LUMO}	μ	C_v
ID Tasks	902	869	945	938	946	931
OOD Tasks	781	885	941	889	886	860

7.2 Maximum atomic forces



Supplementary Figure 14: Histograms of maximum atomic forces for PropMolFlow-generated structures in the ID tasks (filled) and OOD tasks (unfilled). The molecules are directly generated raw molecules after the multi-step filtering. Numbers of remaining molecules can be found in Supplementary Table 14.

We summarize the maximum atomic forces for structures that survive the molecule filtering in Supplementary Figure 14 for both OOD and ID tasks, respectively. Maximum atomic forces imply that most structures should be regarded unrelaxed, where typical maximum atomic forces for relaxed structures should be no larger than 0.05 eV/Å.

Comparing ID and OOD tasks, maximum atomic forces are overall larger for OOD-generated structures versus those in the ID tasks.

8 Additional Details

8.1 QM9 molecular property definitions

- α (Polarizability): Tendency of a molecule to acquire an electric dipole moment when subjected to an external electric field.
- $\Delta\epsilon$: The energy gap between HOMO and LUMO.
- ϵ_{HOMO}: Highest occupied molecule orbital energy.
- ϵ_{LUMO}: Lowest unoccupied molecule orbital energy.
- μ: Dipole moment, which measures the separation of positive and negative charges within a molecule.

- C_v: Heat capacity at room temperature 298.15 K.

8.2 Hyperparameters for training EquiFM conditional models

EquiFM models were trained using flow matching parameterized by an EGNN architecture with 9 layers and 192 hidden features, with a batch size of 64. A deterministic dequantization and a vanilla variance-preserving (VP) probability path for discrete molecular modalities were employed. The model was trained with 2500 epochs, and model inference used 210 time steps with an Euler integration method.

References

[1] Cheng Zeng, Jirui Jin, and Mingjie Liu. Propmolflow data: Version v3. Zenodo. https://doi.org/10.5281/zenodo.17726328, 2025.

[2] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

[3] Noel M O'Boyle, Markus Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3:33, 2011.

[4] Niklas W. A. Gebauer, Michael Gastegger, Stefaan S. P. Hessmann, Klaus-Robert Müller, and Kristof T. Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nat Commun*, 13(1):973, February 2022.

[5] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport, March 2024.

[6] Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant Flow Matching with Hybrid Probability Transport for 3D Molecule Generation. *Advances in Neural Information Processing Systems*, 36:549–568, December 2023.

[7] Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching, November 2023.

[8] Emiel Hoogeboom, Victor Garcia Satorras, Clement Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8867–8887. PMLR, June 2022.

[9] Minkai Xu, Alexander S Powers, Ron O. Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3D molecule generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38592–38610. PMLR, 23–29 Jul 2023.

[10] Ian Dunn and David Ryan Koes. Mixed continuous and categorical flow matching for 3d de novo molecule generation. *CoRR*, abs/2404.19739, 2024.

[11] Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom Blundell, Pietro Lio, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based Drug Design with Equivariant Diffusion Models, September 2024.

[12] Alex Morehead and Jianlin Cheng. Geometry-complete diffusion for 3D molecule generation and optimization. *Commun Chem*, 7(1):1–11, July 2024.

[13] Greg Landrum et al. Rdkit: open-source cheminformatics software. RDKit, 2016.