EnvGeoBench: A Benchmark and MM Adapter for QM-Region Geometry Generation

Anonymous authors
Paper under double-blind review

Abstract

In multiscale molecular simulations of complex chemical systems, the geometry of a chemically active quantum-mechanical (QM) region is shaped by its surrounding environment, typically represented using molecular mechanics (MM). Existing geometry-generation methods largely treat the QM system in isolation, without a systematic approach to incorporating information from surrounding MM regions. This motivates a benchmark beyond isolated-molecule geometry generation that evaluates both structural accuracy and the benefit of environmental information. In this work, we introduce EnvGeoBench, a benchmark for environment-conditioned QM geometry generation from QM/MM simulation records. It defines four tasks for generating QM conformations with decreasing amounts of available reaction information: reaction endpoint conformations and a requested path stage, reaction endpoint conformations alone, a reaction coordinate value alone, or no per-record reaction information. For each task, we specify model inputs, hold out groups of related simulation records, and evaluate generated geometries against their reference structures. For systematic incorporation of MM information and performance evaluation, we introduce a reusable MM adapter that injects a compact representation of the local environment into the QM geometry model, with implementations for a conditional flow model and direct coordinate predictors. We compare four MM conditions to assess the benefit of including the spatial organization and record correspondence of the MM environment: the correct corresponding environment (matched-MM), no environment information (no-MM), randomized atomic directions (shuffled-MM), and an environment from another record (mismatched-MM). Across the four tasks, matched-MM lowers the baseline models' reconstruction error relative to no-MM, with a median relative reduction of 14.3%. In the Window-Coordinate and Environment-Only tasks, it also outperforms shuffled-MM and mismatched-MM. We further evaluate the adapter with flow and diffusion generators to assess its transferability across generative-model families; matched-MM again yields the lowest mean reconstruction error among the four conditions for both families. EnvGeoBench enables systematic quantification of how molecular environments improve QM geometry generation, and its reusable adapter provides a general mechanism for integrating MM context into existing geometry generators.

1 Introduction

Modeling reactions in complex molecular systems requires knowledge about local structures and their surroundings. As an example, in enzymes, the arrangement of catalytic residues and substrates determines which interactions can stabilize a reactive configuration; studies of serine hydrolases connect these conformational preferences to substrate stabilization and catalytic barriers (Surpeta et al., 2022). QM/MM simulations capture this conformation-environment coupling by explicitly modeling a chemically active quantum-mechanical (QM) region within a larger molecular-mechanical (MM) environment (Karelina & Kulik, 2017; Vennelakanti et al., 2022). This hybrid-resolution modeling approach gives rise to conditional QM conformation generation: given the available chemical conditions as priors and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

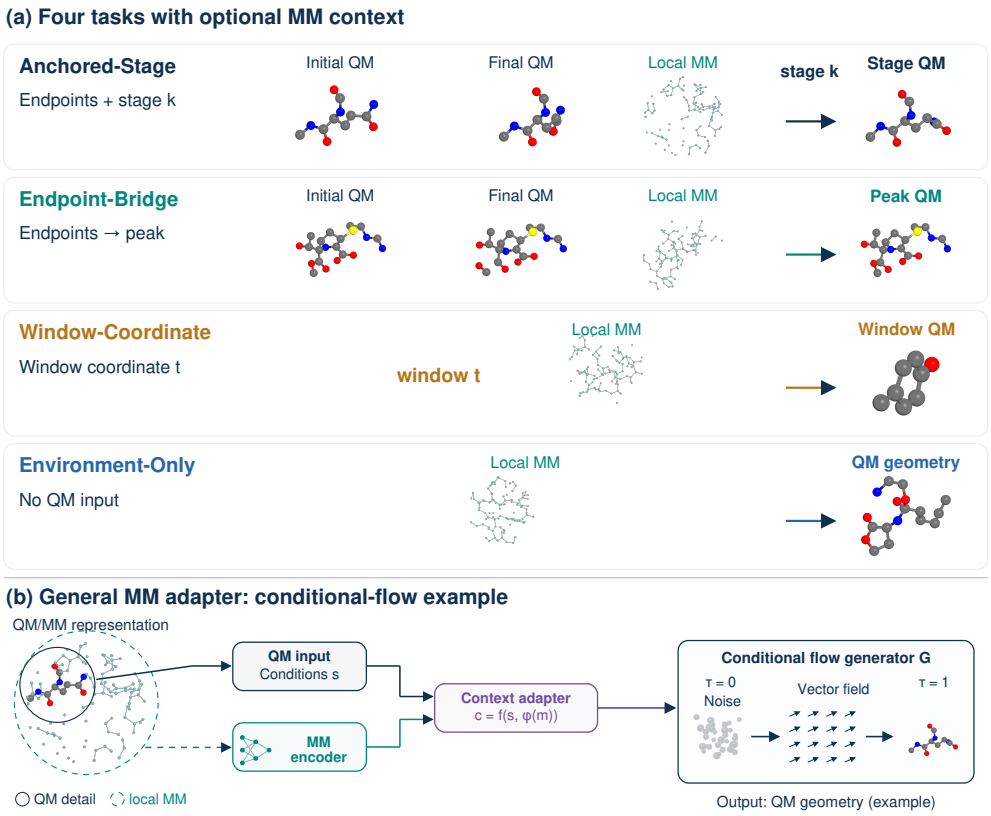

Figure 1: EnvGeoBench and MM conditioning. (a) Four tasks specify what reaction information (such as endpoint geometries or a requested path stage) is available; MM context is optional in every task. Archived QM structures are the targets used to evaluate generated geometries. Renderings show selected QM fragments and local-MM subsets from training records. (b) A conditional-flow example of the general MM adapter combines the available reaction information s with an encoding $\varphi(m)$ of the local MM environment m to form the generator condition $c = f(s, \varphi(m))$. Flow time τ indexes generation steps; the reaction coordinate t specifies a restrained simulation ensemble. T1, T3, and T4 baseline models instead use direct predictors.

the surrounding MM environment, a model needs to generates the three-dimensional coordinates of the QM atoms. Developing reliable models for this task first requires measuring QM-geometry accuracy under each QM/MM information condition. It also requires controlled tests of whether local MM context improves that accuracy, and which environmental information accounts for the improvement.

There have been existing molecular learning methods that address related structural questions. Transition-state generators construct reactive geometries from reactant and product structures (Kim et al., 2024; Duan et al., 2023; 2025). Transition1x and QMrxn20 provide reaction geometries and quantum-chemical labels (Schreiner et al., 2022; von Rudorff et al., 2020), while ATOM3D introduces diverse three-dimensional molecular prediction tasks (Townshend et al., 2021). Docking and pocket-based generation explicitly considers generation of atomic structure within a molecular environment (Corso et al., 2023; Peng et al., 2022). In these cases as well as in our QM/MM generation task, The QM region (or the "chemically interesting region") to be generated is much smaller than its surrounding MM-modeled environment, so a model must extract relevant information from a heterogeneous neighborhood while producing only QM coordinates. However, chemical information known a priori such as reaction endpoint geometries, scalar reaction coordinates, or structural priors usually already somewhat aid accurate reconstruction; the benefit of adding surrounding

MM contexts thus needs to be decoupled from contributions of chemical priors. In this work, we introduce EnvGeoBench, a systematic benchmarking approach to evaluate QM generation within in MM-modeled environment by measuring reconstruction accuracy under four tiers of MM surrounding injectiona: properly supplying the corresponding MM environment (matched-MM), removing it (no-MM), randomizing its atomic coordinates (shuffled-MM), or replacing it with another related record's environment (mismatched-MM). These four tiers of experiments test what MM context adds beyond the reaction information provided for the task and which properties of the environment account for the improvement.

In terms of the difficulty of the QM-generation tasks, EnvGeoBench organizes QM/MM reconstruction into four tiers with different amount of available prior information about the studied chemical reaction or physical transition: generating a structure at a defined intermediate stage (a "reaction coordinate"), with atomic structures at reaction/transition endpoints, such as the "reactants" and "products", also known (T1: Anchored-Stage); reconstructing an archived path peak (e.g. the "transition state") from known endpoint structuress alone (T2: Endpoint-Bridge); reconstructing a geometry sampled near a specified reaction coordinate without knowledge about endpoint structures (T3: Window-Coordinate), or reconstructing a QM structure from MM context alone without any prior information injected about the reaction/transition (T4: Environment-Only; Figure 1(a)). Our full collection of data are organized into eleven task-specific collections, which we call benchmark subsets. Each task fixes the permitted chemistry prior inputs, reference targets, and groups of related records kept together in a data split, forming its information contract (Table 1).

To make these task contracts comparable across model architectures, we provide a consistent MM information adapter. It encodes the local MM environment m as $\varphi(m)$ and combines it with the available reaction information s through $c = f(s, \varphi(m))$ before coordinate generation (Figure 1(b)); the generator outputs only QM coordinates. Researchers can benefit from varying the environment encoder or coordinate generator while retaining the task inputs, targets, grouped splits, and evaluator. To demonstrate the benchmark, we provide task-specific baseline models: an equivariant endpoint-conditioned flow model and direct coordinate predictors. We also evaluate the adapter with flow and diffusion generators, two complementary families of generative models.

Across the eleven subsets, matched-MM lowers the baseline models' mean reconstruction error by 3.3%–72.3% relative to their no-MM counterparts. In the Window-Coordinate and Environment-Only tests, matched-MM further scores lower mean error compared to shuffled-MM or mismatched-MM. Flow and diffusion generator evaluations extend this comparison to T1, where matched-MM has the lowest three-seed mean error across all four arms for both generator families on both evaluated subsets. Our contributions are:

- We introduce EnvGeoBench, a benchmark for evaluating accuracy of QM-region geometry generation in QM/MM environments and the contribution of local MM context. Its four tasks vary the prior knowledge about modeled reaction/chemical process available at inference, including an Environment-Only(T4) evaluation with 4,553 QM/MM configurations in 1,258 grouped split units.
- We define controlled tests of environment utilization. The matched-MM, no-MM, shuffled-MM, and mismatched-MM conditions test the roles of MM access, spatial organization, and record correspondence in reconstruction accuracy.
- We provide a consistent MM adapter with task-matched baseline implementations. It encodes the local environment as conditioning information while the generator itself produces only QM coordinates.
- We test robustness across multiple model architectures. Matched-MM improves reconstruction across both flow and diffusion generators under the same task inputs and evaluation protocol.

2 Related Work

Our work sits at the intersection of reactive-geometry generation, QM/MM modeling, and environment-conditioned molecular learning. Together, previous works in these areas have

established that molecular geometry can be generated from chemical conditions and that surrounding structure can be informative, but they do not provide a controlled evaluation of how MM context improves QM-region geometry generation.

QM/MM calculations represent a chemically active region within its surrounding MM environment, and their behavior depends on the active-region choice and environmental embedding (Kulik et al., 2016; Karelina & Kulik, 2017; Vennelakanti et al., 2022). Equivariant multiscale learning incorporates both regions into a shared geometric representation (Lei et al., 2024). Equivariant potentials such as NequIP and MACE learn energies and forces from atomic configurations (Batzner et al., 2022; Batatia et al., 2022). Their representations can also serve as components of coordinate-generating models. EnvGeoBench focuses instead on evaluating generators that output a fixed set of QM coordinates under explicit input, target, and split specifications.

Reactive-geometry generators construct reactive structures from reactant and product geometries (Kim et al., 2024; Duan et al., 2023; 2025; Luo et al., 2026). Transition1x and QMrxn20 provide reaction geometries with energy or force labels (Schreiner et al., 2022; von Rudorff et al., 2020). These methods motivate generation based on information about modeled reactions, but generally do not test what explicit MM context adds beyond that information. EnvGeoBench retains this generative setting while pairing QM targets with local MM environments.

Environment-conditioned molecular learning shows how surrounding structure can guide a geometric prediction: EquiBind and DiffDock place a supplied ligand in a receptor (Stärk et al., 2022; Corso et al., 2023), while Pocket2Mol and TargetDiff generate ligand structures conditioned on a protein pocket (Peng et al., 2022; Guan et al., 2023). ATOM3D organizes diverse molecular prediction tasks with shared processing and evaluation tools (Townshend et al., 2021). EnvGeoBench brings a comparable evaluation structure to QM/MM geometry generation, fixing the output to QM atom identities and coordinates while varying available reaction information and MM conditions.

3 EnvGeoBench and MM conditioning

EnvGeoBench evaluates how the MM environment informs QM geometry generation under four levels of available prior information about modeled reaction/physical transition.

Across all tasks, data instances are paired QM/MM records along reaction/transition trajectories. QM coordinates are the prediction target; surrounding protein, solvent, or cofactor atoms provide optional MM context. We denote the reaction information available for a record by s, the local MM environment by m, and the archived target by $x^{\star}_{\mathrm{QM}}$. A model generates QM coordinates from the permitted reaction information inputs based on MM environment coniditioning, with sampling noise z for stochastic generators. The target is available to the training loss and, during evaluation, only to the scorer.

Each task specifies the permitted inputs, public structural information available to all models, and the paired target. Public structural information include the output atom identities and ordering, topology, coordinate frame, and source selection rules. Each task also defines related records that should remain together in the same split, so that records from the same simulation replicas, reaction-path lineages, restrained windows, or consecutive frame blocks are not separated in different splits and cause data leakage. These rules form its information contract (Appendix A).

3.1 Four tasks for evaluating environment utilization

Table 1 summarizes the four tiers of information available, illustrated in Figure 1(a). Each benchmark subset is a source data collection evaluated under one task, identified by a task prefix and source shorthand. For example, T1-C06 and T4-C06 share a source data collection but have distinct task-specific records and grouped splits. Appendix A maps these IDs to the source archives and reports dataset sizes and split units.

Table 1: EnvGeoBench tasks and constituent benchmark subsets. Each subset is a collection of simulation records evaluated under one task. All MM input conditions retain the same fixed structural information, such as atom ordering and topology; no-MM removes only explicit MM context. Source mappings and split units are listed in Appendix A, Table 4.

Task	Available reaction information	Target geometry	MAE metric	Subsets
T1: Anchored-Stage	Endpoints + stage	Requested path stage	Heavy-atom pair distance	T1-C06, T1-C08
T2: Endpoint-Bridge	Endpoints	Deposited path peak	Heavy-atom pair distance	T2-KPC, T2-GES
T3: Window-Coordinate	Reaction coordinate	Sampled window record	Aligned Cartesian	T3-L6, T3-L12
T4: Environment-Only	None per record	Current record	Declared-slot pair distance	T4-C06, T4-C08, T4-PGA, T4-HC7, T4-MHE

The tasks are designed to investigate how local MM resolves geometry as the available reaction information is progressively reduced. Anchored-Stage (T1) supplies a discrete stage k in addition to reaction endpoint geometries in the QM region and requests its configuration; Endpoint-Bridge (T2) supplies the endpoints and requests an archived interior "path peak", roughly equivalent to an intermedite state with highest energy/free energy. Window-Coordinate (T3) supplies a reaction coordinate t alone. Umbrella sampling (Torrie & Valleau, 1977) collects correlated frames near this scalar value, so an ensemble of records share s but can have different environments and target geometries. Environment-Only (T4) measures reconstruction from MM context and public structural priors without per-record reaction information. Each subset has a declared MM-neighborhood construction rule. Appendix C gives these rules and grouped splits for all five subsets.

3.2 A general MM adapter

To make the task contracts transferrable between model families, we provide a consistent MM adapter. It treats the available MM environment as conditioning information for the generated output about QM-region geometry: an MM encoder maps m to a compact representation $\varphi(m)$, while the generator produces only QM coordinates. The adapter combines this representation with the available reaction information and the stochastic noise before QM-region coordinate generation:

$$x_{\hat{\mathrm{QM}}} = G_\theta\big(s, \varphi(m); z\big). \tag{1}$$

As mentioned above, for T1–T3, s contains the reaction information available for each record; in T4, s is absent. Figure 1(b) illustrates the adapter for a conditional flow.

The adapter is implemented with different task-specific predictors. T1 uses a direct stage decoder. T2 uses an equivariant endpoint-conditioned flow: a GotenNet backbone (Aykent & Xia, 2025) and vector head transport noisy endpoint-midpoint coordinates to a QM structure, conditioned on relative local-MM geometry. T3 combines a coordinate-conditioned base predictor with an MM-dependent ridge correction. T4 uses an MM slot-attention decoder with a shared encoder and separate output heads for the five subsets. TS-DFM and I2SB adaptations test the adapter with additional flow and diffusion generators. Appendix B.1 gives baseline-model equations, coordinate conventions, and training objectives.

3.3 Evaluation protocol

Accuracy of QM-region reconstruction alone does not identify the contribution of MM context. A lower reconstruction error within the MM environment could be due to access to environmental atoms in general, their spatial organization, or their correspondence to the target record. We therefore compare four MM input conditions, or "arms": matched-MM (the record's corresponding environment), no-MM (MM removed while retaining all available

reaction information and public priors), shuffled-MM (atom directions from the original MM centroid independently resampled while retaining their distances to the centroid, element, and charge), and mismatched-MM (another record's complete local environment). Comparisons within a model family measure the contribution of different aspects of MM contexts; comparisons between methods under the same MM-condition measure model performance.

Within each task and generator, different MM-conditioning arms share the output slots, model budget, optimizer, training schedule, grouped split, and random seed set. Each arm is independently trained and validated. The T1 and T2 tasks compare matched-MM with no-MM; T3, T4, and the TS-DFM and I2SB adaptations on T1 compare all four arms. We report mean ± sample standard deviation across three training or fitting runs. Model comparisons follow each task's fixed sampling protocol (Appendix B.1).

Our primary error metric is pair-distance mean absolute error (D-MAE). It is invariant to global rotation and translation, and in our cases it covers heavy-atom pairs in T1 and T2 and declared QM slots, excluding padding, in T4. T3 uses Cartesian MAE in a shared aligned frame. For the aPGA Anchored-Stage collections, we also evaluate the prediction error measured specifically in the source-defined reaction coordinate, a linear combination of interatomic distances (LCOD). Confidence intervals resample complete held-out groups; variation across model seeds is reported separately. Appendix B.1 specifies metrics, bootstrap procedures, and intervention rules.

Hyperparameters are selected on a validation partition of complete simulation groups and frozen before test evaluation. Input–target separation, split records, and per-seed results are documented in Appendix I.

4 Results

4.1 Matched-MM improves reconstruction across the four tasks

We first study whether local MM improves QM-geometry reconstruction when the reaction information available for the task and public priors are held fixed. Table 2 reports the experiment results of all four tasks. Anchored-Stage and Endpoint-Bridge compare supplying versus omitting MM; Window-Coordinate and Environment-Only also test shuffled and mismatched contexts. Subset IDs follow Table 1.

We also compare the endpoint-conditioned baseline models with the trivial baseline of Cartesian endpoint interpolation: the requested-stage fraction for T1 and the fixed midpoint for T2. These give D-MAEs of 0.4765/0.4419 on T1-C06/C08 and 0.1081/0.1077 on T2-KPC/GES, whereas the matched-MM baseline models give 0.3345/0.2979 and 0.0843/0.0936, respectively (Appendix E.1).

The matched-MM arm has lower mean error than no-MM in all eleven subsets. The reductions range from 3.3% to 72.3% relative to each subset's corresponding no-MM baseline model (Appendix E). In Window-Coordinate and Environment-Only tasks, it also has the lowest mean error among all four MM-conditioning arms. Environment-Only evaluates five subsets drawn from 4,553 QM/MM configurations, with 879 held-out test records. These aggregate accuracy gains establish that MM context improves reconstruction both when reaction information is available for each record and when it is not. We next ask which properties of the environment produce these gains, whether they extend to chemically relevant geometry, and whether the adapter transfers across flow-based and diffusion-based generators.

4.2 Matched-MM context provides record-specific structural information

The aggregate results show that matched-MM conditioning lowers reconstruction error. We next study whether these gains reflect the benefit of record-specific information rather than generic access to an environment, whether they depend on the MM environment's spatial arrangement, and whether they directly improve the interatomic distances used to define the original reaction coordinate, a common key metric of interest when studying chemical reactions.

Table 2: Benchmark results under the task-specific baseline models. Values report task-specific MAE in Å (metrics in Table 1), as mean ± sample standard deviation over three runs; lower is better. The no-MM arm retains the same reaction information and task priors as matched-MM. Red bold marks the lowest mean within each row; a dash denotes an arm not evaluated in this comparison.

Subset ID	no-MM ↓	matched-MM ↓	shuffled-MM ↓	mismatched-MM ↓
T1: Anchored-Stage				
T1-C06	0.3663 ± 0.0007	**0.3345** ± 0.0010	—	—
T1-C08	0.3367 ± 0.0010	**0.2979** ± 0.0003	—	—
T2: Endpoint-Bridge				
T2-KPC	0.1016 ± 0.0037	**0.0843** ± 0.0023	—	—
T2-GES	0.1015 ± 0.0019	**0.0936** ± 0.0021	—	—
T3: Window-Coordinate				
T3-L6	0.3422 ± 0.0035	**0.3300** ± 0.0039	0.3431 ± 0.0035	0.3425 ± 0.0035
T3-L12	0.2982 ± 0.0019	**0.2882** ± 0.0016	0.2985 ± 0.0021	0.2995 ± 0.0018
T4: Environment-Only				
T4-C06	1.1341 ± 0.1787	**0.8451** ± 0.0391	0.8583 ± 0.0308	1.2808 ± 0.1956
T4-C08	1.1197 ± 0.1656	**0.9423** ± 0.1319	1.1113 ± 0.0353	1.0672 ± 0.1138
T4-PGA	2.4019 ± 0.1917	**0.6662** ± 0.0426	1.1150 ± 0.0683	2.2574 ± 0.0245
T4-HC7	0.4991 ± 0.0107	**0.4275** ± 0.0069	0.5265 ± 0.0152	0.5026 ± 0.0194
T4-MHE	0.6746 ± 0.0316	**0.2301** ± 0.0033	0.3695 ± 0.0207	0.7427 ± 0.0196
T4 aggregate	1.1219 ± 0.0244	**0.4923** ± 0.0241	0.6843 ± 0.0131	1.1293 ± 0.0247

In T3 and T4, shuffled-MM randomizes atoms' directions about the original MM centroid while retaining each atom's radius, element, and charge. Each arm is independently trained or fitted and validated. The mismatched-MM arm supplies a donor context under the task-specific rule. The T4 aggregate weights all 879 test records equally. Scores from tasks with different metrics are reported separately. Paired intervals and seed variation are in Appendix E.

Window-Coordinate provides the most direct test of structural ambiguity beyond the available reaction information. In each collection, eight held-out windows contain five records with the same reaction-coordinate value but distinct MM contexts and QM geometries. Internal QM pair distances differ by 0.394–0.546 Å across the records, so the shared coordinate alone does not determine the target geometry. With the same baseline model, matched-MM reduces mean aligned Cartesian MAE relative to no-MM in every window; the collection means improve by 3.6% and 3.3% (Appendix F). These results show that the matched MM environment supplies information about the target geometry beyond the shared reaction-coordinate input.

The four-arm Window-Coordinate comparison then identifies which properties of the MM environment matter. The matched-MM arm has lower mean error than no-MM, shuffled-MM, and mismatched-MM in both held-out collections (Table 2). Relative to no-MM, this result shows that access to MM context can improve reconstruction; its advantage over shuffled and mismatched MM inputs shows that the gain depends on both spatial organization and correspondence to the record. An independent Anchored-Stage temporal-offset control tests correspondence without changing replica identity: for each T1 test record, we retain the endpoint geometries of the QM region and stage request but replace the MM input with the other archived stage from the same replica. With the same three frozen checkpoints, this substitution raises mean D-MAE from 0.3345 to 0.3431 on C06 and from 0.2979 to 0.3078 on C08 (Appendix E.1). This modest but consistent deterioration supports the need for the correctly matched local MM context beyond replica identity.

Environment-Only asks whether record-specific MM information remains useful when no per-record reaction information is supplied. The no-MM arm receives no per-record MM input, whereas the matched-MM arm receives the record's local environment. Across all five subsets, matched-MM has the lowest mean error and improves on no-MM by 629.6 mÅ

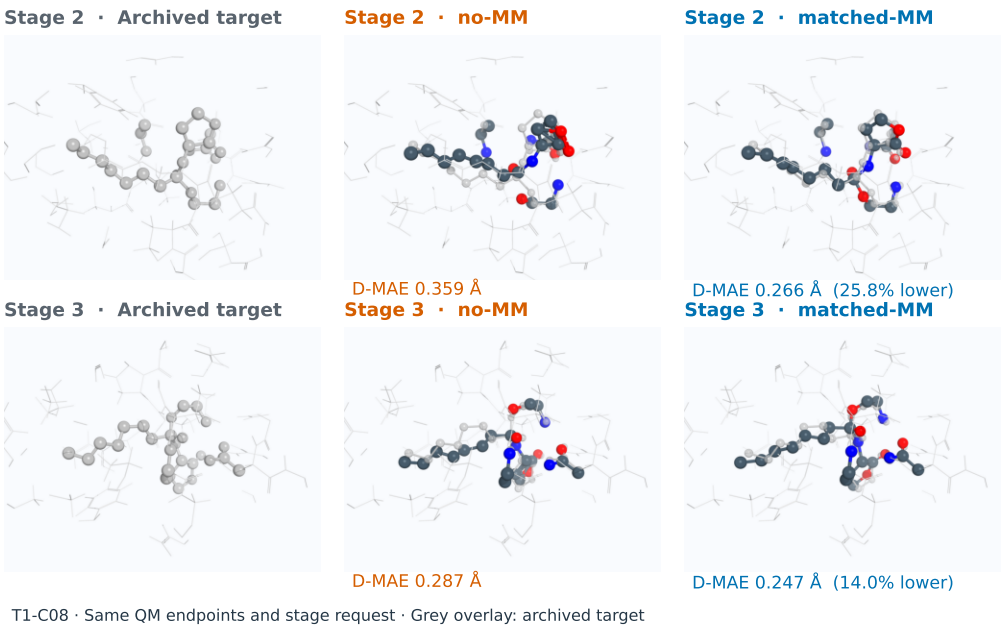

Figure 2: MM conditioning beyond QM endpoints and stage requests. The Anchored-Stage C08 direct stage decoder reconstructs held-out replica 767 at two stages. Matched-MM and no-MM are independently trained seed-101 models receiving the same endpoints and stage request. Each row shares a viewpoint; predictions are colored and archived targets gray. Labels report heavy-atom D-MAE and its reduction relative to no-MM. The replica was selected post hoc for illustration.

(56.1%) in the record-weighted aggregate; it also remains best when the subsets are weighted equally (Appendix C). The shuffled-MM test further shows that access to identity and distaces of individual neighboring MM atoms alone is insufficient: angular shuffling increases mean D-MAE relative to matched-MM by 13.1–448.8 mÅ (1.6%–67.4%) across the five subsets (Table 2). The size of this effect varies by system, but the comparison shows that the neighboring MM atoms' directional arrangement can carry information needed for reconstruction.

We finally test whether lower global geometry error is accompanied by improved accuracy on a reaction-relevant structural quantity. For the prediction of source-defined LCOD which is the linear combination of distances used to define the reaction coordinate, the matched-MM arm reduces MAE from 0.3737 to 0.2754 Å on T1-C06 and from 0.6202 to 0.3565 Å on T1-C08, representing reductions of 26.3% and 42.5%, respectively (Appendix G). The gains therefore extend to the distances used to define progress along the source reaction coordinate.

Figure 2 provides an illustrative example from a held-out Anchored-Stage T1-C08 replica. Both arms receive the same endpoint geometries and stage number. Adding MM reduces D-MAE from 0.359 to 0.266 Å at Stage 2 and from 0.287 to 0.247 Å at Stage 3. At Stage 2, the substrate O10–Asn241β ND2 separation is 3.62 Å in the target, 4.64 Å without MM, and 3.79 Å with matched MM. We note that this replica was selected post hoc for the purpose of best illustration, and this case should not be perceived as establishing expected contact accuracy.

4.3 Training coverage

We next examine training coverage in Window-Coordinate. Increasing the final fitting set from 14 to 34 complete windows reduces matched-MM MAE from 0.3836 to 0.3300 Å on T3-L6 and from 0.3368 to 0.2882 Å on T3-L12. The matched-MM arm has lower mean

error than no-MM at all four dataset sizes in both collections. Appendix H gives the fitting protocol and results for all four dataset sizes.

4.4 Generalization across flow and diffusion generators

The MM adapter is intended to condition generators beyond the task-matched baseline models. We therefore integrate it with the TS-DFM flow generator (Luo et al., 2026) and the I2SB diffusion generator (Liu et al., 2023), and compare the four MM conditions within each model. Appendix D gives implementation and training details.

Table 3 reports four independently trained arms on each generator and Anchored-Stage collection. The matched-MM arm has the lowest three-seed mean test D-MAE for both generators on C06 and C08. Relative to no-MM, it reduces error by 18.7 and 12.7 mÅ for TS-DFM and by 28.5 and 13.8 mÅ for I2SB on C06 and C08, respectively. Both generators also improve on shuffled-MM and mismatched-MM in both collections. Thus matched-MM conditioning with our MM adapter extends to both flow and diffusion generators under the same task specification.

Table 3: Generalization of the adapter across TS-DFM flow and I2SB diffusion generators. Values are held-out heavy-atom D-MAE (Å, mean ± sample standard deviation over three seeds) for four independently trained MM conditions. n denotes held-out replicas; red bold marks the lowest mean in each row.

Generator / Anchored-Stage subset	no-MM (↓)	matched-MM (↓)	shuffled-MM (↓)	mismatched-MM (↓)
TS-DFM / C06 ($n = 42$)	0.3925 ± 0.0117	**0.3739** ± 0.0077	0.3884 ± 0.0065	0.4095 ± 0.0104
TS-DFM / C08 ($n = 48$)	0.3410 ± 0.0078	**0.3283** ± 0.0043	0.3570 ± 0.0057	0.3496 ± 0.0060
I2SB / C06 ($n = 42$)	0.3682 ± 0.0074	**0.3397** ± 0.0052	0.3643 ± 0.0066	0.3693 ± 0.0042
I2SB / C08 ($n = 48$)	0.3179 ± 0.0036	**0.3041** ± 0.0013	0.3155 ± 0.0012	0.3169 ± 0.0047

5 Conclusion

EnvGeoBench provides a common benchmark for QM-region geometry generation and for evaluating the contribution of the MM environment. Its four tasks span different levels of available reaction information, with explicit input contracts, grouped splits, and task-specific evaluators. Across all eleven subsets, matched-MM lowers mean reconstruction error relative to no-MM. The four-arm MM conditioning comparisons show that both spatial organization and record correspondence matter for environment utilization. Further analyses show that MM context helps resolve geometry beyond a shared reaction-coordinate value and improves prediction of source-defined reaction coordinates. Matched-MM baseline models outperform endpoint interpolation on all four T1/T2 subsets; same-replica temporal-offset controls further support the value of record-matched environments in T1.

Alongside the benchmark, we provide a general MM adapter that encodes the MM environment as conditioning information while generating only QM coordinates. Task-matched implementations and integrations with flow and diffusion generators demonstrate its use across model families. On both evaluated T1 subsets, matched-MM achieves the lowest three-seed mean error among all four arms for both the TS-DFM flow and I2SB diffusion generators. Together, the benchmark and adapter provide a common basis for developing environment encoders and QM-coordinate generators and measuring how effectively they use the MM environment.

Reproducibility statement

The EnvGeoBench package provides processed data, grouped splits, evaluators, configurations, selected checkpoints, per-seed scores, checksums, and training and evaluation code; its documentation records source archives and third-party terms needed to reproduce the reported results.

References

Sarp Aykent and Tian Xia. GotenNet: Rethinking efficient 3d equivariant graph neural networks. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=5wxCQDtbMo.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Advances in Neural Information Processing Systems, volume 35, pp. 11423–11436. Curran Associates, Inc., 2022. doi: 10.52202/068431-0830. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4a36c3c51af11ed9f34615b81edb5bbc-Abstract-Conference.html.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature Communications, 13:2453, 2022. doi: 10.1038/s41467-022-29939-5. URL https://doi.org/10.1038/s41467-022-29939-5.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=kKF8_K-mBbS.

Chenru Duan, Yuanqi Du, Haojun Jia, and Heather J. Kulik. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. Nature Computational Science, 3:1045–1055, 2023. doi: 10.1038/s43588-023-00563-7. URL https://doi.org/10.1038/s43588-023-00563-7.

Chenru Duan, Guan-Horng Liu, Yuanqi Du, Tianrong Chen, Qiyuan Zhao, Haojun Jia, Carla P. Gomes, Evangelos A. Theodorou, and Heather J. Kulik. Optimal transport for generating transition states in chemical reactions. Nature Machine Intelligence, 7 (4):615–626, 2025. doi: 10.1038/s42256-025-01010-0. URL https://doi.org/10.1038/s42256-025-01010-0.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=kJqXEPXMsE0.

Michele Invernizzi and Michele Parrinello. Rethinking metadynamics: From bias potentials to probability distributions. The Journal of Physical Chemistry Letters, 11(7):2731–2736, 2020. doi: 10.1021/acs.jpclett.0c00497. URL https://doi.org/10.1021/acs.jpclett.0c00497.

Maria Karelina and Heather J. Kulik. Systematic quantum mechanical region determination in qm/mm simulation. Journal of Chemical Theory and Computation, 13(2):563–576, 2017. doi: 10.1021/acs.jctc.6b01049. URL https://doi.org/10.1021/acs.jctc.6b01049.

Seonghwan Kim, Jeheon Woo, and Woo Youn Kim. Diffusion-based generative ai for exploring transition states from 2d molecular graphs. Nature Communications, 15(1):341, 2024. doi: 10.1038/s41467-023-44629-6. URL https://doi.org/10.1038/s41467-023-44629-6.

Heather J. Kulik, Jianyu Zhang, Judith P. Klinman, and Todd J. Martinez. How large should the qm region be in qm/mm calculations? the case of catechol o-methyltransferase. The Journal of Physical Chemistry B, 120(44):11381–11394, 2016. doi: 10.1021/acs.jpcb.6b07814. URL https://doi.org/10.1021/acs.jpcb.6b07814.

Yao-Kun Lei, Kiyoshi Yagi, and Yuji Sugita. Learning qm/mm potential using equivariant multiscale model. The Journal of Chemical Physics, 160:214109, 2024. doi: 10.1063/5.0205123. URL https://doi.org/10.1063/5.0205123.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image schrödinger bridge. In Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 22042–22062, 2023. URL https://proceedings.mlr.press/v202/liu23ai.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yufei Luo, Xiang Gu, and Jian Sun. Generative flow model on distance geometry for predicting transition states of chemical reactions. Nature Communications, 17(1):7497, 2026. doi: 10.1038/s41467-026-74101-0. URL https://doi.org/10.1038/s41467-026-74101-0.

Yoshiki Narimatsu, Christian Büll, Víctor Taleb, Qinghua Liao, Ismael Compañón, David Sánchez-Navarro, Fabien Durbesson, Renaud Vincentelli, Francisco Corzana, Carme Rovira, Bernard Henrissat, Henrik Clausen, Hiren Joshi, and Ramon Hurtado-Guerrero. Dataset for A Family of Di-Glutamate Mucin Degrading Enzymes that Bridges Glycan Hydrolases and Peptidases, 2023. URL https://doi.org/10.5281/zenodo.10103923.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 17644–17655, 2022. URL https://proceedings.mlr.press/v162/peng22b.html.

Daniel Platero-Rochart. Molecular dynamics and steered QM/MM molecular dynamics simulations files, 2024. URL https://doi.org/10.5281/zenodo.13149215.

Daniel Platero-Rochart, Tatyana Krivobokova, Michael Gastegger, Gilbert Reibnegger, and Pedro A. Sánchez-Murcia. Prediction of enzyme catalysis by computing reaction energy barriers via steered QM/MM molecular dynamics simulations and machine learning. Journal of Chemical Information and Modeling, 63(15):4623–4632, 2023. doi: 10.1021/acs.jcim.3c00772. URL https://doi.org/10.1021/acs.jcim.3c00772.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 9323–9332, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x – a dataset for building generalizable reactive machine learning potentials. Scientific Data, 9:779, 2022. doi: 10.1038/s41597-022-01870-w. URL https://doi.org/10.1038/s41597-022-01870-w.

Luka Simsive, Oleksii Zdorevskyi, and Vivek Sharma. Proton transfer through a charged conduit in respiratory complex I: Long-range effects and conformational gating. Journal of Chemical Information and Modeling, 65(19):10600–10612, 2025a. doi: 10.1021/acs.jcim.5c01365. URL https://doi.org/10.1021/acs.jcim.5c01365.

Luka Simsive, Oleksii Zdorevskyi, and Vivek Sharma. Proton transfer through a charged conduit in respiratory complex I – long range effects and conformational gating, 2025b. URL https://doi.org/10.5281/zenodo.15747940.

Zilin Song and Peng Tao. Dataset: 1,000 QM/MM minimum energy pathway conformations for the deacylation reactions of GES-5/imipenem, 2022. URL https://doi.org/10.5281/zenodo.6617839.

Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. EquiBind: Geometric deep learning for drug binding structure prediction. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 20503–20521, 2022. URL https://proceedings.mlr.press/v162/stark22b.html.

Bartlomiej Surpeta and Jan Brezovsky. Common dynamic determinants govern quorum quenching activity in N-terminal serine hydrolases, 2022. URL https://doi.org/10.5281/zenodo.6424939.

Bartlomiej Surpeta, Michal Grulich, Andrea Palyzova, Helena Maresova, and Jan Brezovsky. Common dynamic determinants govern quorum quenching activity in N-terminal serine hydrolases. ACS Catalysis, 12:6359–6374, 2022. doi: 10.1021/acscatal.2c00569. URL https://doi.org/10.1021/acscatal.2c00569.

G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. Journal of Computational Physics, 23(2):187–199, 1977. doi: 10.1016/0021-9991(77)90121-8. URL https://doi.org/10.1016/0021-9991(77)90121-8.

Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. ATOM3D: Tasks on molecules in three dimensions. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c45147dee729311ef5b5c3003946c48f-Abstract-round1.html.

Vyshnavi Vennelakanti, Azadeh Nazemi, Rimsha Mehmood, Adam H. Steeves, and Heather J. Kulik. Harder, better, faster, stronger: large-scale qm and qm/mm for predictive modeling in enzymes and proteins. Current Opinion in Structural Biology, 72:9–17, 2022. doi: 10.1016/j.sbi.2021.07.004. URL https://doi.org/10.1016/j.sbi.2021.07.004.

Guido Falk von Rudorff, Stefan N. Heinen, Marco Bragato, and O. Anatole von Lilienfeld. Thousands of reactants and transition states for competing e2 and sn2 reactions. Machine Learning: Science and Technology, 1(4):045026, 2020. doi: 10.1088/2632-2153/aba822. URL https://doi.org/10.1088/2632-2153/aba822.

Chao Yin. 800 QM/MM minimum energy pathway conformations for the deacylation reactions of KPC-2/imipenem, 2022. URL https://doi.org/10.5281/zenodo.7387266.

A Task contracts and archive mapping

We formalize each task's information contract as $C = (\mathcal{S}, \mathcal{K}, \mathcal{T}, \mathcal{G}, \mathcal{M})$: per-record inputs, public task-level priors, the reference target, the group kept intact when splitting the data, and the metric. The public priors $\mathcal{K}$ declare fixed output atom identities and ordering (the atom-slot schema), molecular connectivity (topology), coordinate frame/origin, and source selection and ordering rules. Model architectures can vary while C remains fixed. T1 and T2 expose the declared QM atom identities explicitly through qm_z. T4 instead retains these identities in its public output-slot schema without supplying a per-record QM atom-type tensor to the model.

Table 4 maps the constituent subsets in main-text Table 1 to their original source collections and split units.

The main paper uses benchmark-defined task names. Table 5 gives the contract, split unit, and evaluator target; the released data card maps these names to the source archives and immutable record manifests.

Each public task declares its benchmark baseline model and permitted source fields, public task-level priors (slot schema/topology, frame, and preprocessing), a fixed identifier for each

group of related simulation records, and the evaluation metric. The loader returns the permitted inputs and a record ID. A submission returns that ID, the ordered QM output slots, and an $N_{\mathrm{QM}} \times 3$ coordinate array. The evaluator matches predictions to reference structures by record ID, checks shape, atom order, and finite coordinates, and aggregates errors over the declared simulation groups. In particular, anchored-stage sources contain qm_z, qm_stage1_xyz, qm_stage4_xyz, and stage_request; the matched-MM arm additionally receives mm_xyz, mm_z, and mm_charge.

An Anchored-Stage record requires that the two QM endpoints, requested stage, and contemporaneous MM frame belong to the same replica. Endpoint-Bridge additionally requires an archived interior peak and a complete path lineage; Window-Coordinate requires a complete restrained window. Anchored-Stage draws its two source collections from the same enzyme archive (Surpeta & Brezovsky, 2022): T1-C06 contains 420 records from 210 replicas, split into 294/42/84 records for train/validation/test; T1-C08 contains 480 records from 240 replicas, split into 336/48/96. Both frames belonging to one replica remain in the same partition. Environment-Only combines five subsets (aPGA C06/C08, ecPGA, HC7, and MHETase), totaling 4,553 records in 1,258 split groups and split into 3,313/361/879 train/validation/test records at group boundaries. T4-C06/C08 and T4-PGA use the aPGA and ecPGA collections, respectively, in the enzyme archive (Surpeta & Brezovsky, 2022; Surpeta et al., 2022); T4-HC7 uses the HC7 archive (Narimatsu et al., 2023). T4-MHE uses the MHETase simulation archive (Platero-Rochart, 2024; Platero-Rochart et al., 2023). The coordinate tasks draw source files from two collections in a respiratory-complex umbrella-sampling archive (Simsive et al., 2025b;a). EnvGeoBench processing retains 42 complete umbrella windows per collection; eight windows (40 records) are held out and the remaining 34 windows (170 records) support training and inner validation. The endpoint source files correspond to KPC-2/imipenem and GES-5/imipenem (Yin, 2022; Song & Tao, 2022). T2-KPC contains 800 path records: 200 for each combination of wild-type or F72Y KPC-2 with the imipenem Δ_1 or Δ_2 tautomer. T2-GES contains 1,000 path records, 500 for each of the two imipenem tautomers. These combinations define the four KPC and two GES source conditions. Records sharing a lineage ID across conditions stay in the same partition: 636/164 outer-training/test records for KPC and 814/186 for GES, with 41 and 93 held-out lineage groups, respectively.

B Data selection and baseline-model details

Every primary configuration is selected on an inner validation partition before outer testing. The bond- and angle-loss weights for the Anchored-Stage baseline model are selected on T1-

Table 4: Benchmark subset IDs, original source collections, and held-out units. IDs distinguish task-specific evaluation subsets; subsets from the same source can have different records and splits.

Subset ID	Source collection	Held-out unit	Test groups
T1-C06	aPGA-C06	Replica	42
T1-C08	aPGA-C08	Replica	48
T2-KPC	KPC-2	Lineage	41
T2-GES	GES-5	Lineage	93
T3-L6	Complex-I/L6	Complete window	8
T3-L12	Complex-I/L12	Complete window	8
T4-C06	aPGA-C06	Replica	44
T4-C08	aPGA-C08	Replica	48
T4-PGA	ecPGA	Replica	96
T4-HC7	HC7	Frame block	41
T4-MHE	MHETase	Frame block	19

T1-C06/C08 contain 420/480 records. T3-L6/L12 each contain 210 records in 42 windows. T4 contains 4,553 records in 1,258 groups, split into 3,313/361/879 train/validation/test records. Shared sources across tasks have task-specific records and splits. Source citations and full partitions are in Appendix A.

Table 5: Task contracts in EnvGeoBench. At test time, target structures are available only to the evaluator. Submitted models must follow the specified inputs, outputs, and split rules; their architecture can differ from the benchmark baseline model.

Task	permitted QM input	MM frame at inference	evaluator-only target / metric	baseline model
T1: Anchored-Stage	QM atom types, two endpoints, stage request	same-record MM frame	archived QM stage / heavy-atom D-MAE	direct stage decoder
T2: Endpoint-Bridge	QM atom types, two endpoints	paired path MM record	deposited peak / heavy-atom D-MAE	endpoint-conditioned flow
T3: Window-Coordinate	umbrella coordinate	same-record MM frame	aligned QM geometry / Cartesian MAE	coordinate predictor + MM ridge correction
T4: Environment-Only	none per record	current-frame local MM environment	archived QM geometry / declared-slot D-MAE	slot-attention decoder

C06 and transferred to T1-C08. All four Environment-Only arms share the same architecture and optimizer settings.

Anchored-Stage baseline models use up to 512 nearest MM atoms. The context-free arm has the same decoder and parameter count but masks the complete MM tensor. Window-Coordinate baseline models combine a predictor based on Fourier features of the reaction coordinate with a ridge residual that includes an intercept. The 512 MM atoms nearest the fixed deposited reference QM centroid are retained in topology order. Each arm independently fits its feature scaler and selects its ridge penalty on inner validation before refitting on the full training partition. The Environment-Only baseline model embeds MM coordinates, radial distance, elements, and charges with a shared encoder. Learned queries for each subset's output QM atoms attend to these embeddings and emit coordinates in the declared QM atom order.

B.1 Baseline-model equations and metric definitions

The T1 direct stage decoder predicts a relative-vector residual around the endpoint midpoint. Let $x_i^{(1)}$ and $x_i^{(4)}$ denote the two endpoint positions, $\bar{x}^{(r)}$ their centroids, and m_j an MM position. It predicts

$$\hat{x}_i = \frac{x_i^{(1)} + x_i^{(4)}}{2} + \sum_{r\in\{1,4\}} a_i^{(r)}(x_i^{(r)} - \bar{x}^{(r)}) + \sum_{\ell} g_{i\ell}\frac{1}{M}\sum_{j=1}^{M} w_{ij\ell}(m_j - x_i^{(1)}). \tag{2}$$

The learned scalar weights $w_{ij\ell}$ and gates $a_i^{(r)}, g_{i\ell}$ use endpoint pair distances, QM slot and element embeddings, the requested stage, MM elements and charges, and MM–endpoint distances. Index ℓ spans six vector channels and M counts non-padding MM atoms; the no-MM arm sets the MM contribution to zero. The direct stage decoder's topology-regularized objective is

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_b\mathcal{L}_{\text{bond}} + \lambda_a\mathcal{L}_{\text{angle}}. \tag{3}$$

The bond and angle penalties supplement the reconstruction loss, with weights λ_b and λ_a. All quantities and weights are selected from training and inner-validation data before outer testing.

For the T3 Window-Coordinate baseline model, the aligned output is

$$x_{\hat{\text{QM}}} = f_{\text{base}}(\text{Fourier}(t)) + W\,\text{vec}(\widetilde{m}_1, \ldots, \widetilde{m}_{M_{\text{T3}}}) + b, \qquad M_{\text{T3}} = 512, \tag{4}$$

where $\widetilde{m}$ denotes MM coordinates standardized using training-only feature means and scales; W and b are the fitted ridge coefficients and intercept; vec flattens the MM coordinates into a feature vector.

Table 6: Geometric preprocessing used to construct model inputs. "Reference" denotes the fixed deposited structure or topology used to define a task.

Contract	Selection center and membership	Ordering	Coordinate frame and provenance
Anchored-stage	midpoint of the two visible endpoint centroids; MM atoms within 15 Å, then nearest 512	stable increasing source-only radius	archive frame; the direct stage decoder uses MM–endpoint relative vectors
Endpoint-bridge	midpoint of visible endpoint centroids; common endpoint shell within 8 Å, capped at 512	stable minimum endpoint-shell radius; paired endpoint slots preserve archive order	joint endpoint archive frame; all transforms use permitted endpoints
Window-coordinate	512 MM atoms nearest the fixed deposited reference QM centroid	fixed PSF topology order	each frame translated by the mean of 512 fixed reference-nearest MM identities; membership and identities fixed before record targets are read
T4-C06/C08, T4-PGA	512 nearest MM atoms to a topology-defined, non-QM scaffold anchor	stable current-frame distance to the MM-only origin	periodic minimum-image coordinates about fixed MM anchor atoms
T4-MHE	512 nearest MM atoms to fixed non-QM anchor residues	stable increasing distance to anchor mean	translated by the current-frame anchor mean
T4-HC7	512 nearest MM atoms to the archived QM centroid in each record	increasing Euclidean distance to the QM centroid	deposited Cartesian frame; selection is performed during dataset construction

For T4 Environment-Only, a learned query u_i for each output QM slot attends to MM keys k_j and values v_j:

$$\alpha_{ij} = \mathrm{softmax}_j(u_i^\top k_j/\sqrt{d}), \qquad c_i = \sum_j \alpha_{ij} v_j, \qquad b_i = \sum_j \alpha_{ij} m_j, \tag{5}$$

where d is the embedding width. Its head predicts

$$\hat{x}_i = b_i + \mathrm{MLP}_{\mathrm{cell}}([c_i, u_i]). \tag{6}$$

Here $\mathrm{MLP}_{\mathrm{cell}}$ is the output head for that subset. The encoder and heads are trained jointly on the training partition. The test-time input contains only mm_xyz, mm_z, and mm_charge.

For the D-MAE tasks, the evaluator reports

$$\text{D-MAE}(x_{\hat{\mathrm{QM}}}, x^\star_{\mathrm{QM}}) = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left| \|\hat{x}_i - \hat{x}_j\|_2 - \|x^\star_i - x^\star_j\|_2 \right|, \tag{7}$$

over the declared labeled atom pairs $\mathcal{P}$: heavy-atom pairs for Anchored-Stage and Endpoint-Bridge, and all pairs among the non-padding QM slots for Environment-Only, including hydrogen where present. The field reactive_mask identifies all non-padding QM slots. Environment-Only first computes D-MAE within each record and then averages records equally within each subset or across the full test set; mean and sample standard deviation are computed across the three runs. Window-Coordinate uses Cartesian MAE in its deposited aligned frame. For aPGA, the supplementary LCOD readout is $q_{\mathrm{LCOD}}(x) = -\|x_8 - x_9\|_2 + \|x_1 - x_9\|_2$ and reports $|q_{\mathrm{LCOD}}(x_{\hat{\mathrm{QM}}}) - q_{\mathrm{LCOD}}(x^\star_{\mathrm{QM}})|$. Confidence intervals use paired bootstrap draws over the task's indivisible held-out groups after averaging the frozen training seeds. The reported intervals use 20,000 draws: complete replicas for Anchored-Stage and complete lineage groups for Endpoint-Bridge, retaining all conditions of each sampled lineage. Geometry-error reductions are reported as control minus matched-MM (no-MM for the primary comparison), so positive values favor matched-MM. Relative reductions follow Appendix E.

Table 6 specifies each task's MM selection, ordering, and coordinate frame.

The primary baseline models use three training or fitting runs. Within each study, all compared arms share the same seed set. All four Window-Coordinate arms use the same eight

complete held-out windows. KPC/GES endpoint evaluation additionally averages errors over five fixed rollout noise realizations within each record, then averages records within each source condition and conditions equally. The paired bootstrap resamples complete lineage groups across conditions. Seed values are specified in the code configuration. T2 uses an endpoint-conditioned velocity model with Gaussian noise added to the endpoint midpoint and 20 Euler updates. The score averages the geometry errors of the five individual samples; it uses neither best-sample selection nor coordinate averaging.

C Environment-Only details

T4-HC7 construction. We use the 711 selected configurations in the deposited HC7 QM/MM OPES (Invernizzi & Parrinello, 2020) path (Narimatsu et al., 2023). The archive's fixed 28-atom QM selection defines the target slots. For each configuration, we compute their centroid, exclude the QM atoms from the full system, and retain the nearest 512 MM atoms, ordered by increasing Euclidean distance to that centroid. Coordinates remain in the deposited Cartesian frame; element identities and partial charges follow the matched topology. The exported source stores only mm_xyz, mm_z, and mm_charge, while QM coordinates are stored separately for training losses and evaluation. The pair-distance metric is invariant to translation, so recovering the QM centroid alone cannot reduce the reported error. A QM-independent anchor would test the remaining contribution of QM-based neighborhood selection. Each block contains three consecutive records in the selected path. Hashing block IDs prefixed by the subset identifier assigns 173/23/41 whole blocks (519/69/123 records) to train/validation/test. The other subsets retain replica groups (T4-C06/C08 and T4-PGA) or 20-record temporal blocks (T4-MHE); the released manifests fix every assignment.

Baseline model and evaluation. Environment-Only uses an MM-only input schema: local coordinates, elements, and charges. The five subsets have different fixed QM atom counts and orderings, so the baseline model uses a separate output head for each subset and a shared 192-hidden-unit context interface. All arms use AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-3}, weight decay 10^{-5}, batch size 32, gradient clipping at 1.0, and an 80-epoch budget; we retain the checkpoint with the lowest mean D-MAE over validation batches. Test scores average per-record errors from these fixed checkpoints. Three independently initialized fits use this same recipe and a shared seed set for every arm. Shuffled MM preserves each atom's centroid-relative radius, element identity, and charge while resampling its direction; the mismatched-MM arm replaces the complete local MM environment with one from another record. Aggregate and per-seed baseline scores are provided in results_main.json and reference_scores.json in the public package.

Aggregation across T4 subsets. Record weighting measures error over the pooled test collection; subset-equal weighting measures the average of its five constituent subset errors. For each fixed training seed, we first compute the five per-subset means and then average them, retaining the seed pairing across arms. Table 7 reports both summaries. The matched-MM arm reduces no-MM error by 56.1% with record weighting and 46.6% with subset-equal weighting. It has the lowest mean in each individual subset as well as under both aggregate definitions.

Table 7: T4 aggregation sensitivity. Values are declared-slot D-MAE in Å (mean ± sample standard deviation over three training seeds). Subset weighting averages the five subset errors equally within each seed; record weighting gives each of the 879 test records equal weight.

Arm	Record-weighted	Subset-equal
no-MM	1.1219 ± 0.0244	1.1659 ± 0.0495
matched-MM	0.4923 ± 0.0241	0.6222 ± 0.0372
shuffled-MM	0.6843 ± 0.0131	0.7961 ± 0.0089
mismatched-MM	1.1293 ± 0.0247	1.1701 ± 0.0429

D External generator portability

TS-DFM flow. Section 4.4 and Table 3 include the adaptation of the published TS-DFM distance-geometry flow core (Luo et al., 2026). The Anchored-Stage adapter supplies endpoint and stage conditions alongside the coordinates and elements of at most 16 local MM atoms. Each of the four arms is independently initialized and trained for 120 epochs with seeds 101, 202, and 303. All arms share the selected configuration (hidden width 64, two layers, learning rate 3×10^{-4}) and select their own checkpoint by validation reconstruction D-MAE, evaluated every five epochs. The flow and reconstruction use 10 and 25 steps, respectively.

The shuffled-MM arm follows the T4 isotropic angular rule: it preserves each selected atom's radial distance about the selected-MM centroid and independently draws its direction uniformly on the sphere, preserving atom identities. Mismatched MM uses another record from the same partition and stage as the donor. The test protocol and configurations are fixed before test evaluation. TS-DFM final evaluations run on CPU and use the same 42/48 held-out replicas as the T1 comparison. Paired intervals use 20,000 complete-replica bootstrap draws of seed-averaged record differences. The adapted graph, masks, reconstruction constraints, and conditioner are part of the evaluated implementation. Table 8 reports these intervals for all three controls. Per-seed baseline scores and the selected weights are indexed in Appendix I.

Table 8: TS-DFM paired heavy-atom D-MAE reductions (mÅ), control minus matched-MM, with 95% intervals from 20,000 complete-replica bootstrap draws after averaging three seeds. Positive values favor matched-MM.

Subset	Control	Reduction [95% CI]
T1-C06	no-MM	18.68 [11.79, 25.56]
	shuffled-MM	14.55 [7.46, 21.84]
	mismatched-MM	35.65 [27.07, 44.24]
T1-C08	no-MM	12.70 [8.35, 17.39]
	shuffled-MM	28.77 [22.08, 35.67]
	mismatched-MM	21.29 [16.03, 26.78]

I2SB diffusion. We adapt I2SB (Liu et al., 2023) to molecular coordinates using its native stochastic forward bridge and posterior sampler with an EGNN denoiser (Satorras et al., 2021) and the same spatial MM conditioner used by TS-DFM. The bridge connects the source endpoint-midpoint prior to the target QM geometry during training; generation uses only the source endpoints, requested stage, and permitted MM context. The task, replica partitions, selection of 16 MM atoms, four MM conditions, and heavy-atom D-MAE are the same as in the TS-DFM experiment. The mismatched-MM arm uses a fixed cyclic offset of 17 within the same partition and stage. The no-MM arm removes the MM nodes and zeroes the MM conditioner while retaining the endpoint QM context.

The diffusion configuration was selected on C08 training/validation data using development seed 505 and then fixed for all confirmation runs on both subsets. Each arm is trained independently for 120 epochs with seeds 101, 202, and 303, learning rate 3×10^{-4} and weight decay 10^{-5}. The denoiser has hidden width 128 and four layers, with joint width 64 and slot width 128; the conditioner has width 16 and cutoff 8 Å. We use 1,000 diffusion intervals, $\beta_{\max} = 0.3$, 20 stochastic sampling steps, EMA decay 0.99, AdamW with gradient clipping at 1, and learning-rate decay 0.99 every 1,000 updates. Each arm selects its own checkpoint by minimum generated validation D-MAE at epochs 20, 40, 60, 80, 100, and 120, choosing the earlier epoch in a tie. TS-DFM and I2SB thus use the same task, input conditions, and training duration, with their respective networks, numerical solvers, and validation schedules. The same diffusion recipe is trained separately on C06 and C08.

The test sets contain 84 records from 42 replicas on C06 and 96 records from 48 replicas on C08. Diffusion sampling runs on A800 GPUs, using the same sampling-seed schedule across arms and training seeds. The public package provides the selected checkpoints and per-seed baseline scores (Appendix I).

Table 9: Task-specific errors and paired relative gains. MAE is reported in Å (mean ± sample standard deviation over three runs); relative reductions include 95% paired group-bootstrap intervals. T1/T2 use heavy-atom D-MAE, T3 aligned Cartesian MAE, and T4 declared-slot D-MAE. Intervals resample groups after seed averaging.

Task	Subset ID	no-MM	matched-MM	Reduction (%)	Seed range (%)
T1	T1-C06	0.3663 ± 0.0007	0.3345 ± 0.0010	8.7 [6.7, 10.6]	[8.2, 9.1]
T1	T1-C08	0.3367 ± 0.0010	0.2979 ± 0.0003	11.5 [10.0, 13.1]	[11.3, 11.8]
T2	T2-KPC	0.1016 ± 0.0037	0.0843 ± 0.0023	17.0 [16.5, 17.5]	[13.4, 18.8]
T2	T2-GES	0.1015 ± 0.0019	0.0936 ± 0.0021	7.7 [7.1, 8.3]	[4.8, 10.6]
T3	T3-L6	0.3422 ± 0.0035	0.3300 ± 0.0039	3.6 [2.2, 5.0]	[3.4, 3.7]
T3	T3-L12	0.2982 ± 0.0019	0.2882 ± 0.0016	3.3 [1.5, 5.6]	[3.3, 3.5]
T4	T4-C06	1.1341 ± 0.1787	0.8451 ± 0.0391	25.5 [18.1, 32.0]	[10.2, 35.1]
T4	T4-C08	1.1197 ± 0.1656	0.9423 ± 0.1319	15.8 [11.3, 20.1]	[−6.1, 37.3]
T4	T4-PGA	2.4019 ± 0.1917	0.6662 ± 0.0426	72.3 [71.4, 73.1]	[70.9, 73.1]
T4	T4-HC7	0.4991 ± 0.0107	0.4275 ± 0.0069	14.3 [9.6, 18.7]	[11.1, 16.0]
T4	T4-MHE	0.6746 ± 0.0316	0.2301 ± 0.0033	65.9 [64.5, 67.2]	[63.6, 67.8]

E Paired effects and fit variation

Table 2 compares the task-specific matched-MM and no-MM baseline models. The reported relative gain is the ratio of their mean errors, $100(1 - \bar{e}_{\text{Matched MM}}/\bar{e}_{\text{No MM}})$, not a mean of per-record ratios. Table 9 reports scores in each task's geometry metric and variation across fitted models.

For each arm, we average per-record errors across the three fixed fits, then resample complete held-out groups with replacement for 20,000 paired bootstrap draws. Every draw uses the same groups in both arms and recomputes the ratio. We divide each resampled error sum by its record count, giving each record equal weight. Endpoint lineages retain their associated conditions; HC7 uses its declared within-path frame blocks. These intervals quantify held-out-group variation conditional on the fitted models, while sample standard deviations and seed-level relative-gain ranges report fit variation.

E.1 Endpoint and temporal controls

Table 10 evaluates two direct controls for the endpoint-conditioned tasks. The endpoint baseline is a literal Cartesian interpolation of the supplied QM endpoints: T1 uses the requested stage fraction, and T2 uses the fixed midpoint. For T2, we report both the midpoint and the five-point grid at fractions $\{0, 0.25, 0.5, 0.75, 1\}$, whose errors are averaged. The deterministic midpoint is repeated five times for the same evaluator, giving exactly its single-prediction error. The temporal control retains every T1 record's endpoints and stage request, while replacing the complete local-MM input with the other archived stage from the same replica. Each snapshot uses the task's fixed per-record local-neighborhood construction, and the replay uses the three frozen matched-MM checkpoints.

Table 10: Endpoint and temporal controls for T1 and T2. Values are held-out heavy-atom D-MAE in Å; lower is better. Endpoint interpolation is a deterministic Cartesian baseline: stage fractions for T1, grid and midpoint reported separately for T2. The temporal-offset values are means ± sample standard deviations over the three frozen T1 checkpoints.

Task/subset	matched-MM	stage/grid interpolation	T2 midpoint	temporal-offset MM
T1-C06	0.3345 ± 0.0010	0.4765	—	0.3431 ± 0.0028
T1-C08	0.2979 ± 0.0003	0.4419	—	0.3078 ± 0.0006
T2-KPC	0.0843 ± 0.0023	0.1403	0.1081	—
T2-GES	0.0936 ± 0.0021	0.1305	0.1077	—

F Same-request window analysis

We evaluate all eight held-out windows in each Window-Coordinate collection. Each window contains five records with the same scalar reaction-coordinate value and distinct 512-slot MM coordinate arrays. For every window, we compute the mean absolute difference over its ten pairs of internal QM distance matrices and the three-run no-MM-minus-matched-MM coordinate-MAE gain. Table 11 reports these values; saved predictions and evaluator inputs are included in the public package. Window labels follow the source archive: $-4, -2, \ldots, 78$ (42 windows), not consecutive indices. The held-out labels are $0, 10, \ldots, 70$.

Table 11: All held-out same-request groups. Geometry difference is the mean between-record internal-distance difference; gain is the reduction in coordinate MAE with matched-MM (mean ± sample standard deviation over three runs).

Collection	Window label	Geometry difference (Å)	Gain (mÅ)
T3-L6	0	0.4627	20.43 ± 0.38
	10	0.4139	16.56 ± 1.20
	20	0.5458	4.86 ± 1.81
	30	0.4603	12.52 ± 1.32
	40	0.4856	5.01 ± 0.91
	50	0.4528	19.26 ± 0.94
	60	0.4283	4.42 ± 0.62
	70	0.5072	14.18 ± 0.47
T3-L12	0	0.3935	25.10 ± 1.66
	10	0.4218	19.51 ± 0.98
	20	0.3976	0.43 ± 0.84
	30	0.4667	7.33 ± 1.09
	40	0.4516	8.34 ± 0.90
	50	0.4399	14.37 ± 1.14
	60	0.4230	1.26 ± 2.57
	70	0.5244	3.40 ± 0.61

G aPGA LCOD evaluation

The deposited steering input for T1-C06 and T1-C08 defines the reaction coordinate as a linear combination of distances (LCOD): $-\|x_8 - x_9\|_2 + \|x_1 - x_9\|_2$, using zero-based QM-slot indices. The corresponding one-based AMBER topology atom indices 3167, 3168, and 3159 map to QM slots 8, 9, and 1, respectively. We apply this fixed expression to all frozen outer-test predictions from the three existing seeds. We compute absolute errors for each seed, average them across seeds, and then average within each complete two-stage replica. The reported interval uses 20,000 paired replica-bootstrap draws for no-MM minus matched-MM. Table 12 reports the errors and paired reductions.

Table 12: aPGA Anchored-Stage reaction-coordinate accuracy. LCOD is the continuous reaction coordinate declared in the source steering input. Parentheses list held-out replicas/records. Lower is better; red bold marks the lower error. Intervals are 95% paired replica-bootstrap intervals for no-MM minus matched-MM; positive reductions favor matched-MM.

Task	no-MM (Å)	matched-MM (Å)	Reduction (Å), 95% CI
T1-C06 (42/84)	0.3737	**0.2754**	0.0983 [0.0180, 0.1844]
T1-C08 (48/96)	0.6202	**0.3565**	0.2637 [0.1499, 0.3779]

The metric measures agreement with the archived reaction-coordinate value.

H Training coverage

Training-window ablation. For each T3 collection, we retain the same seven validation windows and eight test windows. A deterministic farthest-coordinate ordering of the 27 eligible inner-training windows, starting with labels -4 and 78, defines nested sets of 7, 14, 20, and 27 windows. Final refitting includes the seven validation windows, yielding 14, 21, 27, and 34 fitting windows. Each setting uses three seeds (1001, 2003, 3001), a 256-wide Fourier-coordinate base trained for 2,500 epochs, and an independently standardized ridge residual with a nine-value penalty grid from 10^{-2} to 10^{6}. Validation selects the penalty separately for each arm; donor assignment stays within each fitting or evaluation partition.

Table 13 reports all four fitting sizes. Matched-MM has lower mean error than no-MM in every setting. The paired 95% intervals exclude zero except for T3-L6 with 14 fitting windows.

Table 13: Training coverage in Window-Coordinate. Errors are Cartesian MAE in Å (mean ± sample standard deviation over three seeds). Reductions are no-MM minus matched-MM in mÅ, with 95% intervals from 20,000 paired held-out-window bootstrap draws after seed averaging. Fitting counts include the seven validation windows used in final refitting.

Subset	Fitting windows	no-MM	matched-MM	Reduction [95% CI]
T3-L6	14	0.3885 ± 0.0038	0.3836 ± 0.0046	4.84 $[-1.63, 13.32]$
	21	0.3729 ± 0.0078	0.3635 ± 0.0077	9.44 $[2.54, 17.03]$
	27	0.3519 ± 0.0060	0.3405 ± 0.0052	11.33 $[7.32, 15.24]$
	34	0.3422 ± 0.0035	0.3300 ± 0.0039	12.15 $[7.79, 16.39]$
T3-L12	14	0.3437 ± 0.0092	0.3368 ± 0.0086	6.83 $[0.87, 13.70]$
	21	0.3052 ± 0.0029	0.2975 ± 0.0035	7.66 $[2.10, 13.45]$
	27	0.3016 ± 0.0022	0.2940 ± 0.0024	7.57 $[2.59, 13.12]$
	34	0.2982 ± 0.0019	0.2882 ± 0.0016	9.97 $[4.40, 15.98]$

I Data, code, and reproducibility

The public package contains the processed data for all eleven subsets, grouped splits, task evaluators, six baseline model families, training and evaluation entry points, and 90 selected checkpoints. Its data and model specifications are in docs/data.md and docs/models.md; source archives and attribution are listed in docs/attribution.md. The distribution is EnvGeoBench.tar.gz, with SHA-256 9a447cd899004f4c386f85db0e1ffc73000c58508cbbe8dbb73ddf97ab772f69 . File checksums are listed in CHECKSUMS.sha256.

The model names exposed by the training and evaluation commands are stage, endpoint, window, and environment for T1–T4, respectively. The additional T1 generators are distance_flow (TS-DFM) and diffusion (I2SB). results_main.json contains the means and sample standard deviations in Table 2; reference_scores.json contains per-seed baseline scores for all six model families, including those in Table 3. Configurations are in configs/, and checkpoints/manifest.json indexes the selected weights and their hashes.

From the extracted package root, the following commands install the benchmark, run its API tests, and evaluate the T3-L6 baseline model:

```
python -m pip install -e '.[test,baselines]'
python -m pytest tests -q
python examples/evaluate.py --model window --subset T3-L6 --seed 1001
```

examples/train.py provides the corresponding training entry point. The model guide specifies the additional environments for the generative models and endpoint flow. Diffusion sampling and sampling from the endpoint baseline model use CUDA; reproducing their per-seed scores requires the same sampling backend. The TS-DFM core is fetched from a pinned upstream revision with examples/fetch_distance_flow.py. Third-party terms, including the I2SB license, are retained in THIRD_PARTY.md and third_party/.

Environment-Only evaluates all declared QM slots, including hydrogen where present, and excludes padding. The slot counts are 100/106 for T4-C06/C08, 100 or 106 for T4-PGA, 34 for T4-MHE, and 28 for T4-HC7.

J Use of language models

Language-model tools assisted literature retrieval, idea development, experimental workflow and code development, result organization, and manuscript editing. The authors reviewed the outputs and take responsibility for the manuscript, results, and scientific claims.