

25 June 2026

MusicMol: Bidirectional Molecule–Music Translation via Neighborhood-Aligned Generation and Transformer Decoding

Jiayi Tang^{1,2}, Ziyang Zhu¹, Chao Cui¹, Nan Ke^{1,2}, Chenxiao Song¹, Yong Li¹, Yuan Zhou¹, Chunting Yu¹, Pengfei Jing², Chaoyang Zhu^{2,3}, Zaixing Yang^{2,4}, Wanxiang Shen^{1,2}

1. College of Pharmaceutical Sciences, Zhejiang University

2. Future Learning Center, Zhejiang University

3. School of Medicine, Zhejiang University

4. College of Life Sciences, Zhejiang University

Abstract

What if a molecule could be heard as a piano melody, and a melody could be used to generate a molecule? We present MusicMol, a bidirectional framework for generative translation between small molecules and symbolic music. In the molecule-to-music direction, a label-free, rule-guided conditional flow-matching generator trained on two million ChEMBL compounds maps each heavy atom to four note parameters—pitch, inter-onset step, duration, and velocity. To make the generated music chemically meaningful, MusicMol aligns molecular similarity neighborhoods with differentiable piano-roll neighborhoods using a kernelized alignment loss that combines softmax KL-divergence with rank preservation, so that structurally similar molecules are mapped to acoustically similar scores. In the music-to-molecule direction, a REMI/BPE encoder–decoder Transformer autoregressively decodes symbolic music into valid SELFIES molecular strings, enabling user-composed or performed music to generate molecular candidates. To support this bidirectional loop, we construct, to our knowledge, the largest molecule–MIDI paired dataset to date, comprising ten million compound–score pairs. On held-out music–molecule pairs, the inverse model achieves high token-level generation accuracy and near-perfect molecular validity. Across 30 molecular activity prediction tasks, the music-derived piano-roll representation outperforms four established molecular encoders spanning SMILES, graph, image, and 3D modalities in frozen linear-probe Spearman correlation, R2, and RMSE, and achieves the lowest local label incoherence index (LII) from k=5 to k=100. Deployed as an interactive platform at the Future Learning Center of Zhejiang University, MusicMol lets users hear molecules in real time and translate music into

molecular candidates, providing a perceptual and generative interface between chemistry and music.

MusicMol: Bidirectional Molecule–Music Translation via Neighborhood-Aligned Generation and Transformer Decoding

Jiayi Tang^{1,4*}, Ziyang Zhu^{1,*}, Chao Cui^{1,*}, Nan Ke^{1,4}, Chenxiao Song¹, Yong Li¹, Yuan Zhou¹,
Chunting Yu¹, Pengfei Jing⁴, Chaoyang Zhu^{2,4}, Zaixing Yang^{3,4}, and Wanxiang Shen^{1,4,‡}

¹College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China

²School of Medicine, Zhejiang University, Hangzhou, China

³College of Life Sciences, Zhejiang University, Hangzhou, China

⁴Future Learning Center, Zhejiang University, Hangzhou, China

* These authors contributed equally to this work; ‡ Correspondence: shenwx25@zju.edu.cn

What if a molecule could be heard as a piano melody, and a melody could be used to generate a molecule? We present MUSICMOL, a bidirectional framework for generative translation between small molecules and symbolic music. In the molecule-to-music direction, a label-free, rule-guided conditional flow-matching generator trained on two million ChEMBL compounds maps each heavy atom to four note parameters—pitch, inter-onset step, duration, and velocity. To make the generated music chemically meaningful, MUSICMOL aligns molecular similarity neighborhoods with differentiable piano-roll neighborhoods using a kernelized alignment loss that combines softmax KL-divergence with rank preservation, so that structurally similar molecules are mapped to acoustically similar scores. In the music-to-molecule direction, a REMI/BPE encoder–decoder Transformer autoregressively decodes symbolic music into valid SELFIES molecular strings, enabling user-composed or performed music to generate molecular candidates. To support this bidirectional loop, we construct, to our knowledge, the largest molecule–MIDI paired dataset to date, comprising ten million compound–score pairs. On held-out music–molecule pairs, the inverse model achieves high token-level generation accuracy and near-perfect molecular validity. Across 30 molecular activity prediction tasks, the music-derived piano-roll representation outperforms four established molecular encoders spanning SMILES, graph, image, and 3D modalities in frozen linear-probe Spearman correlation, R^2 , and RMSE, and achieves the lowest local label incoherence index (LII) from $k=5$ to $k=100$. Deployed as an interactive platform at the Future Learning Center of Zhejiang University, MUSICMOL lets users hear molecules in real time and translate music into molecular candidates, providing a perceptual and generative interface between chemistry and music.

Introduction

What if a molecule could be heard as a piano melody, and a melody could be translated back into a molecular candidate? Molecular representation is central to chemical machine learning, governing how structural information is encoded and made accessible to learning algorithms across property prediction, de novo design, and reaction modeling¹. Existing modalities—SMILES² and SELFIES³ strings, molecular topological graphs^{4,5}, 2D depiction images⁶, and three-dimensional geometry representations⁷—are precise but perceptually opaque: none can be heard, and none provides an intuitive creative interface through which non-specialists can explore molecular space. Symbolic music offers a different kind of structured representation: pitch, inter-onset step, duration, and velocity are encoded as discrete, machine-readable events, making music simultaneously human-perceptible, computationally tractable, and suitable as a carrier of molecular information.

The feasibility of connecting chemistry and music has been established from several directions. Data sonification has shown that structured scientific information can be rendered acoustically for communication and discovery^{8,9}. Advances in AI for music—from long-range Transformer generators^{10,11} to large-scale symbolic pretraining^{12,13}—demonstrate that symbolic music is a discrete language amenable to modern sequence-learning models. Prior work has explored sonification for chemistry education¹⁴, amino-acid-to-music mappings for protein design¹⁵, and molecule-to-composition transfer as a carrier of chemical information¹⁶. Yet two limitations remain. First, existing chemistry–music mappings do not explicitly align molecular similarity neighborhoods with musical similarity neighborhoods; as a result, generated scores may be perceptually interesting but chemically unorganized. Second, to our knowledge, music has not been systematically used as an input modality for molecular generation, leaving the reverse direction of molecule–music translation largely unexplored.

Here we present MUSICMOL, a bidirectional framework for generative translation between molecules and symbolic music (**Figure 1**). The structure-to-MIDI route (S2M) uses a label-free, rule-guided conditional flow-matching generator trained on two million ChEMBL compounds to assign each heavy atom four note parameters—pitch, inter-onset step, duration, and velocity. To make the generated music chemically meaningful, S2M aligns molecular-fingerprint neighborhoods with differentiable piano-roll neighborhoods through a kernelized alignment loss combining neighborhood-distribution matching and rank preservation. CHEMMUSICRULE provides chemically informed warm-start targets that encode molecular topology as musically coherent note scaffolds, while neighborhood alignment refines the mapping so that structurally similar molecules

produce acoustically similar scores.

The MIDI-to-structure route (M2S) performs the reverse translation. Because large-scale experimental molecule–music pairs do not exist, we first apply the trained S2M route to selected PUBCHEM molecules to generate symbolic scores, and pair each score with the corresponding molecular SELFIES string. This procedure yields, to our knowledge, the largest molecule–MIDI paired dataset to date, comprising ten million compound–score pairs for supervised music-to-molecule learning. A REMI/BPE encoder–decoder Transformer^{11,17,18} then encodes symbolic music and autoregressively decodes it into valid SELFIES molecular strings³, allowing user-composed, performed, or S2M-generated music to be translated into molecular candidates. Because S2M and M2S share a common symbolic music representation, MUSICMOL forms a closed bidirectional loop in which molecules can be rendered as music and music can serve as an input for molecular candidate generation.

Across 30 molecular activity prediction tasks, the music-derived piano-roll representation outperforms four established encoders spanning complementary modalities—SMILES-based ChemBERTa, graph-based GROVER, image-based ImageMol, and 3D-geometry-based Uni-Mol—in frozen linear-probe Spearman correlation, R^2 , and RMSE, and achieves the lowest local incoherence index (LII) at every neighborhood size from $k=5$ to $k=100$. In the reverse direction, the M2S Transformer achieves high token-level generation accuracy and near-perfect molecular validity on held-out music–molecule pairs. Deployed at the Future Learning Center of Zhejiang University as an on-site interactive installation, MUSICMOL lets users hear molecules in real time and translate music into molecular candidates, providing a perceptual and generative interface between chemistry and music.

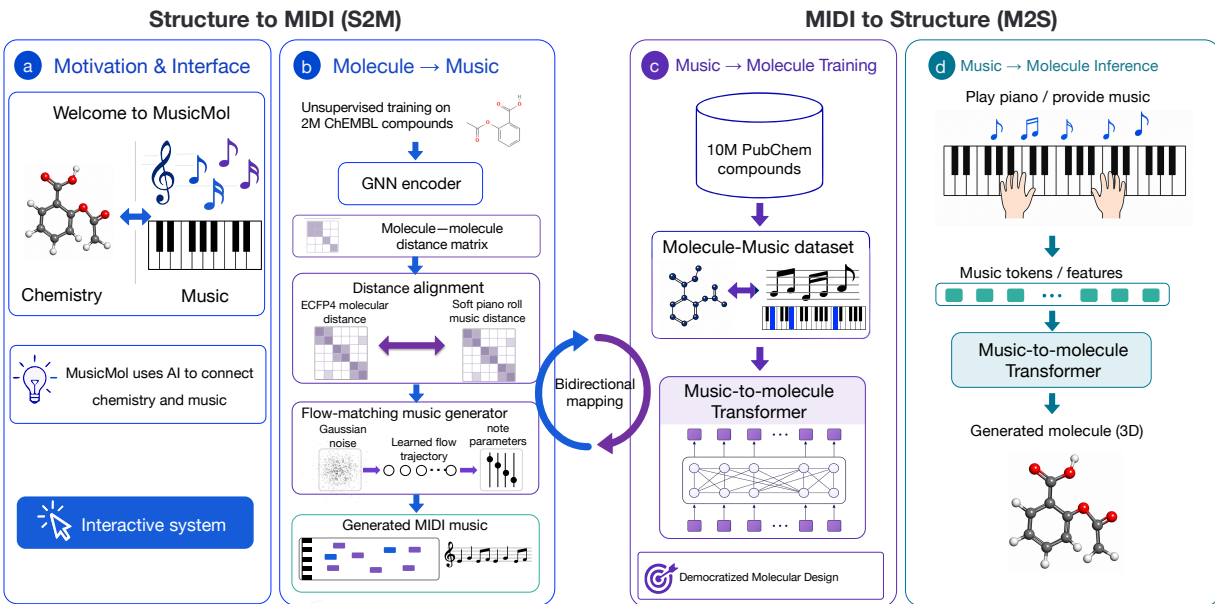


Figure 1: Overview of the MUSICMOL framework. The framework comprises four components. (a) An interactive interface accepts molecule input and returns generated MIDI music in real time, illustrating the perceptual motivation for chemistry–music translation. (b) The structure-to-MIDI (S2M) route encodes each molecule with a GNN to obtain atom-context representations, computes a molecule–molecule distance matrix, and trains a conditional flow-matching generator to produce four-dimensional note parameters; a kernelized distance alignment loss matches the soft piano-roll music distance to the molecular distance, preserving chemical neighborhoods in music space. (c) The MIDI-to-structure (M2S) route pairs ten million PUBCHEM compound–score pairs to train a music-to-molecule Transformer on the resulting molecule–music dataset. (d) At inference, any symbolic music input—played on a piano or otherwise provided—is encoded and decoded into a novel molecular structure with predicted physicochemical properties.

Results

Neighborhood-aligned molecule-to-music generation. MUSICMOL translates molecular structures into symbolic music through an atom-level conditional flow-matching generator (Figure 2). Each heavy atom in the input molecule is encoded into a 256-dimensional context vector by a four-layer residual message-passing neural network, and these atom-context vectors condition a shared vector field $v_{\theta}(z_t, t \mid h_i)$. Starting from atom-level Gaussian noise, the vector field transports each atom toward a four-dimensional note representation consisting of pitch, inter-onset step, duration, and velocity. The resulting atom-level note parameters are then quantized into a symbolic music21 score and exported as MIDI or MusicXML.

To provide a chemically informed initialization, we use CHEMMUSICRULE to define rule-guided warm-start targets before neighborhood alignment refines the mapping. CHEMMUSI-

CRULE maps molecular topology to musically coherent note targets: pitch follows a BFS-depth arch contour from a central atom, inter-onset step reflects ring size and local branching patterns, duration controls articulation, and velocity reflects structural importance, assigning stronger dynamics to ring junctions and branch points and weaker dynamics to terminal atoms. This rule-guided target stabilizes early flow-matching training while leaving the final music representation to be shaped by molecular neighborhood alignment.

The S2M generator is optimized by combining flow matching with a kernelized neighborhood-alignment loss. For each mini-batch, differentiable soft piano-rolls rendered from the generated note parameters define a music distance matrix D_{pred} , while Morgan radius-2 fingerprints define a molecular distance matrix D_{target} . Rather than directly regressing one distance matrix onto the other, both matrices are converted by a softmax kernel into row-stochastic neighborhood probability distributions. A KL-divergence term aligns molecular and music neighborhood distributions, and a rank-preservation hinge encourages the nearest molecular neighbors of each anchor to remain closer in music space than sampled distant molecules. Thus, S2M learns a music representation in which structurally similar molecules are mapped to acoustically similar symbolic scores.

Training on two million ChEMBL compounds showed stable convergence of the neighborhood-aligned S2M generator (**Figure 3**). Total training and validation losses decreased steadily throughout optimization, reaching a best validation loss of 0.129 at epoch 99 (**Figure 3a**). As the alignment weight increased, the validation neighborhood-alignment loss decreased (**Figure 3b,d**), while the batch-wise Pearson correlation between Morgan fingerprint distances and soft piano-roll music distances rose to approximately 0.90 by the end of training (**Figure 3c**). These trends indicate that the generator not only learns to produce valid symbolic note parameters, but also organizes the generated music space according to molecular similarity neighborhoods.

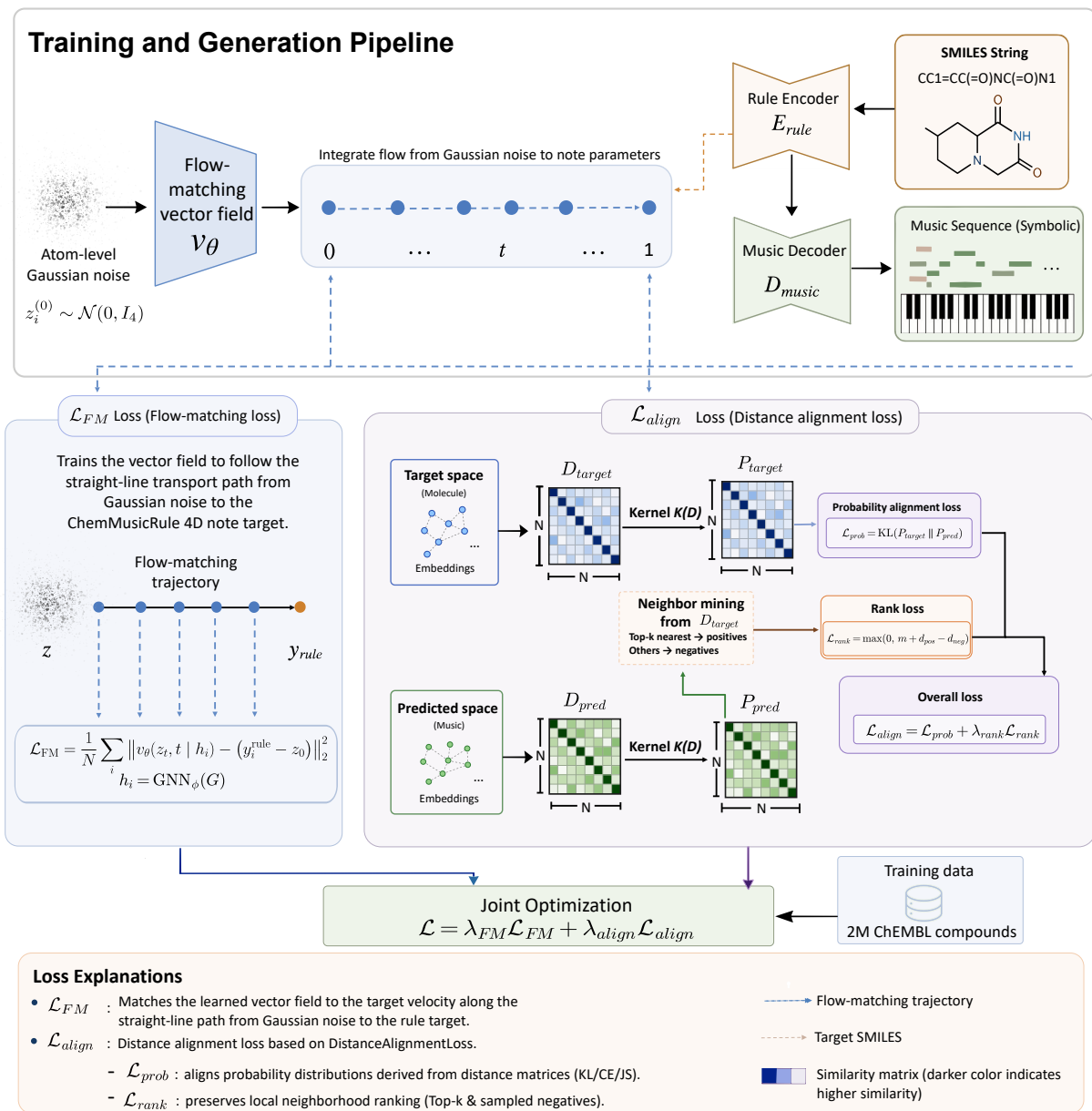


Figure 2: Neighborhood-aligned S2M generation pathway. *Top:* A conditional flow-matching vector field v_θ transports atom-level Gaussian noise $z_i^{(0)} \sim \mathcal{N}(0, I_4)$ toward CHEMMUSICRULE note targets y^{rule} along a straight-line interpolant, conditioned on atom-context representations h_i from the GNN encoder. Integration yields four note parameters—pitch, inter-onset step, duration, and velocity—that are assembled into a symbolic score and rendered as a differentiable soft piano-roll. *Bottom:* The generator is jointly optimized with a flow-matching loss and a neighborhood-alignment loss. The flow-matching loss $\mathcal{L}_{FM}$ regresses the learned vector field onto the target velocity along the straight-line transport path. The alignment loss applies a softmax kernel $K(D)$ to both the molecular distance matrix D_{target} and the music distance matrix D_{pred} , producing neighborhood distributions P_{target} and P_{pred} . A KL-divergence term aligns these distributions, and a rank-preservation hinge enforces that top- k molecular neighbors remain closer in music space than sampled negatives. The joint objective $\mathcal{L} = \lambda_{FM} \mathcal{L}_{FM} + \lambda_{align} \mathcal{L}_{align}$ is minimized end-to-end.

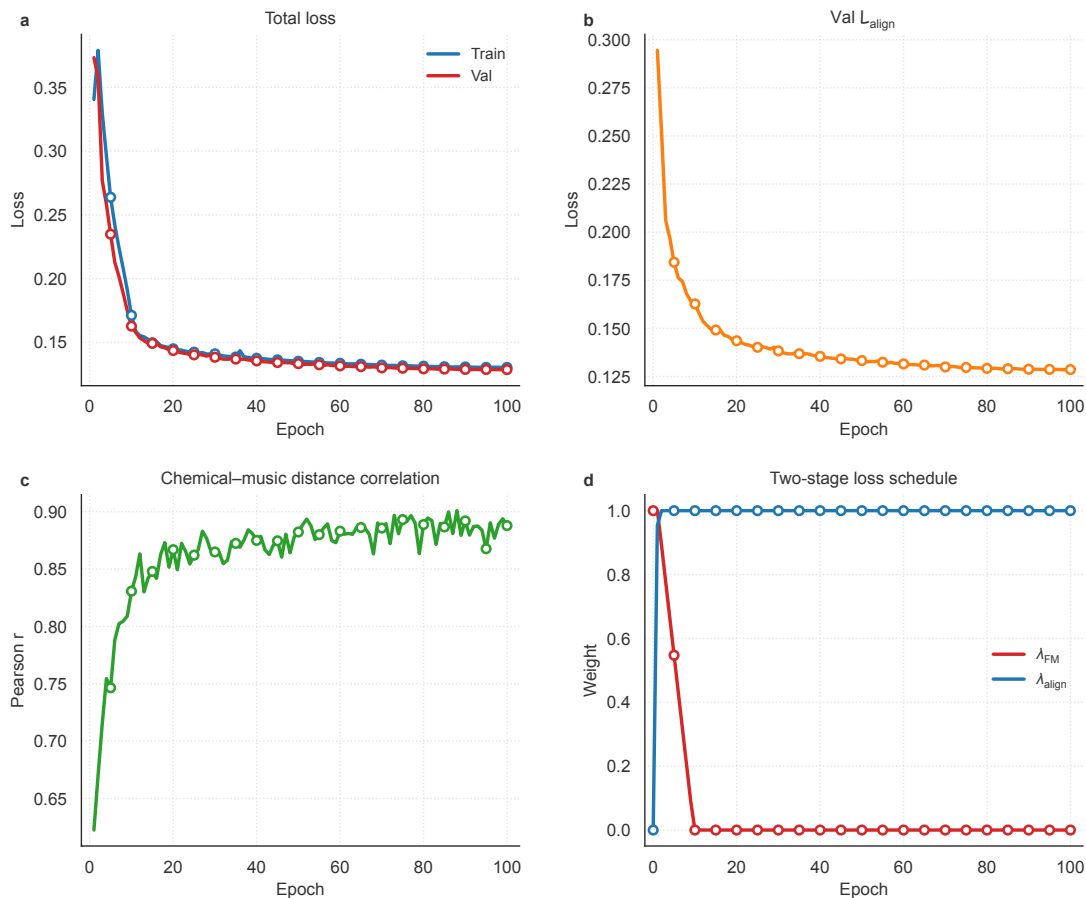


Figure 3: S2M structure-to-MIDI training dynamics. The S2M route encodes molecular graphs with an MPNN and generates four-dimensional note vectors [pitch, inter-onset step, duration, velocity] via conditional flow matching. Training minimizes $\mathcal{L} = \lambda_{FM}\mathcal{L} * FM + \lambda * align\mathcal{L} * align$, where $\mathcal{L} * align$ aligns neighborhood distributions induced by Morgan radius-2 molecular distances and soft piano-roll music distances. **a**, Total training and validation loss; the best validation loss reaches 0.129 at epoch 99. **b**, Validation neighborhood-alignment loss. **c**, Pearson correlation between molecular and music distance matrices during training. **d**, Two-stage loss schedule; λ_{FM} decays to zero by epoch ~ 10 , while λ_{align} reaches 1.0.

Music-to-molecule decoding from S2M-generated paired data. The MIDI-to-structure route (M2S) performs the reverse translation, mapping symbolic music back to molecular sequences (Figure 4). Because large-scale experimental molecule–music pairs do not exist, we first apply the trained S2M route to selected PUBCHEM molecules to generate symbolic scores, and pair each score with the corresponding molecular SELFIES string. This procedure yields, to our knowledge, the largest molecule–MIDI paired dataset to date, comprising ten million compound–score pairs for supervised music-to-molecule learning.

In M2S, a symbolic music file—whether generated by S2M, composed by a user, or performed on a physical piano—is tokenized with a REMI-style event representation¹¹ and compressed with byte-pair encoding¹⁸ into a sequence of music token IDs with a vocabulary of 30,000 tokens. An encoder–decoder Transformer¹⁷ encodes this music sequence into contextual representations and autoregressively decodes it into SELFIES molecular tokens³. The decoded SELFIES sequence is converted into a canonical molecular string that is chemically valid under the SELFIES grammar.

We next evaluated whether the S2M-generated paired corpus supports accurate inverse translation from music to molecular sequences (Figure 5). On held-out music–molecule pairs, the M2S Transformer achieved a raw SELFIES exact-match accuracy of approximately 0.50 and a canonical SMILES exact-match accuracy of approximately 0.50 by epoch 15. Token-level generation accuracy exceeded 0.95, and molecular validity approached 1.0 throughout training. These results show that symbolic music generated by the S2M route contains sufficient information for a Transformer decoder to recover valid molecular sequences, establishing a learnable inverse mapping from music to molecules.

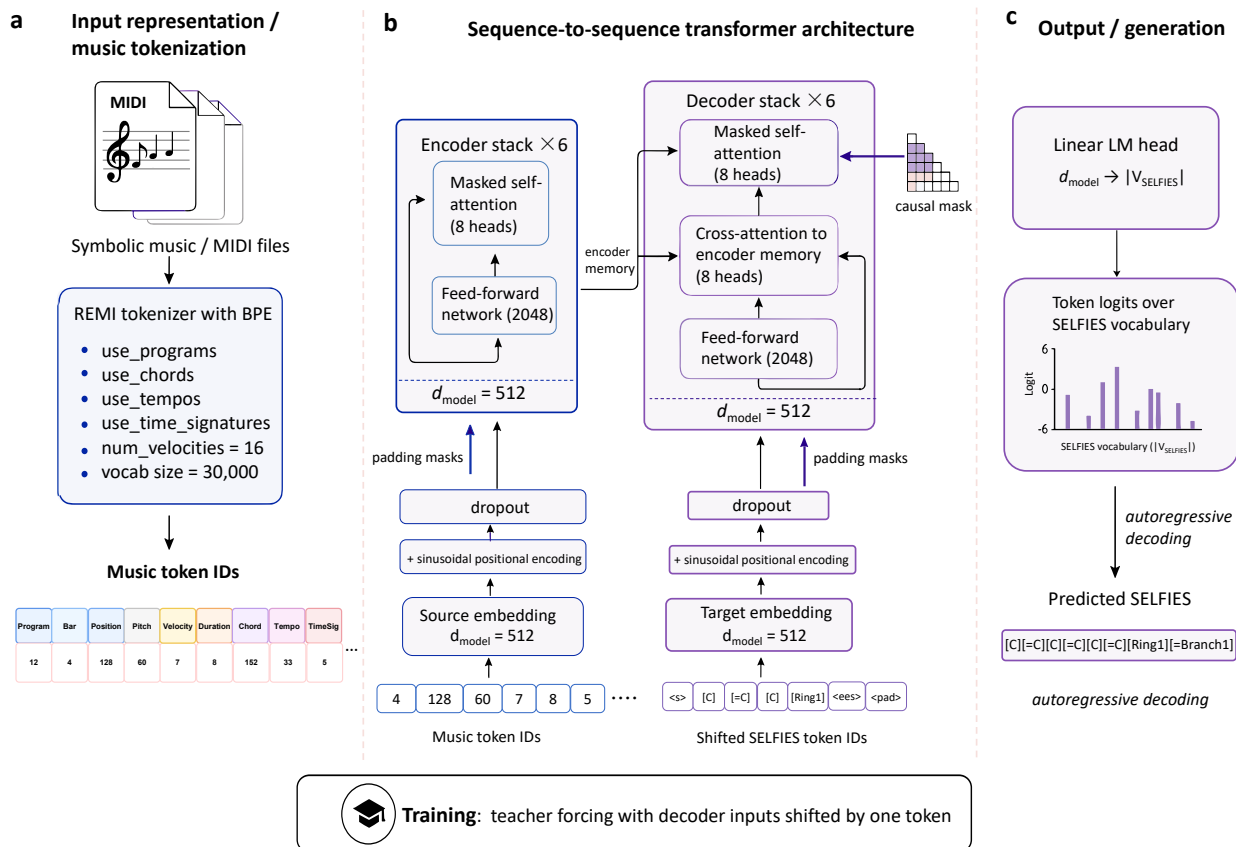


Figure 4: M2S music-to-SELFIES decoding pathway. **a, Input representation and music tokenization.** Symbolic music or MIDI files are converted into an event sequence using a REMI-style tokenizer configured with programs, chords, tempos, time signatures, and 16 velocity levels, then compressed with byte-pair encoding into a compact vocabulary of 30,000 music token IDs. **b, Sequence-to-sequence Transformer architecture.** Music token IDs are embedded into $d_{\text{model}}=512$ dimensional representations and processed by a six-layer encoder. A six-layer decoder attends to the encoder memory via cross-attention and to the shifted SELFIES token sequence via masked self-attention, with a causal mask enforcing autoregressive generation. Training uses teacher forcing with decoder inputs shifted by one token. **c, Output and generation.** A linear language-model head projects decoder outputs to logits over the SELFIES vocabulary; tokens are generated autoregressively until an end-of-sequence token is produced, and the resulting sequence is decoded into a canonical molecular string.

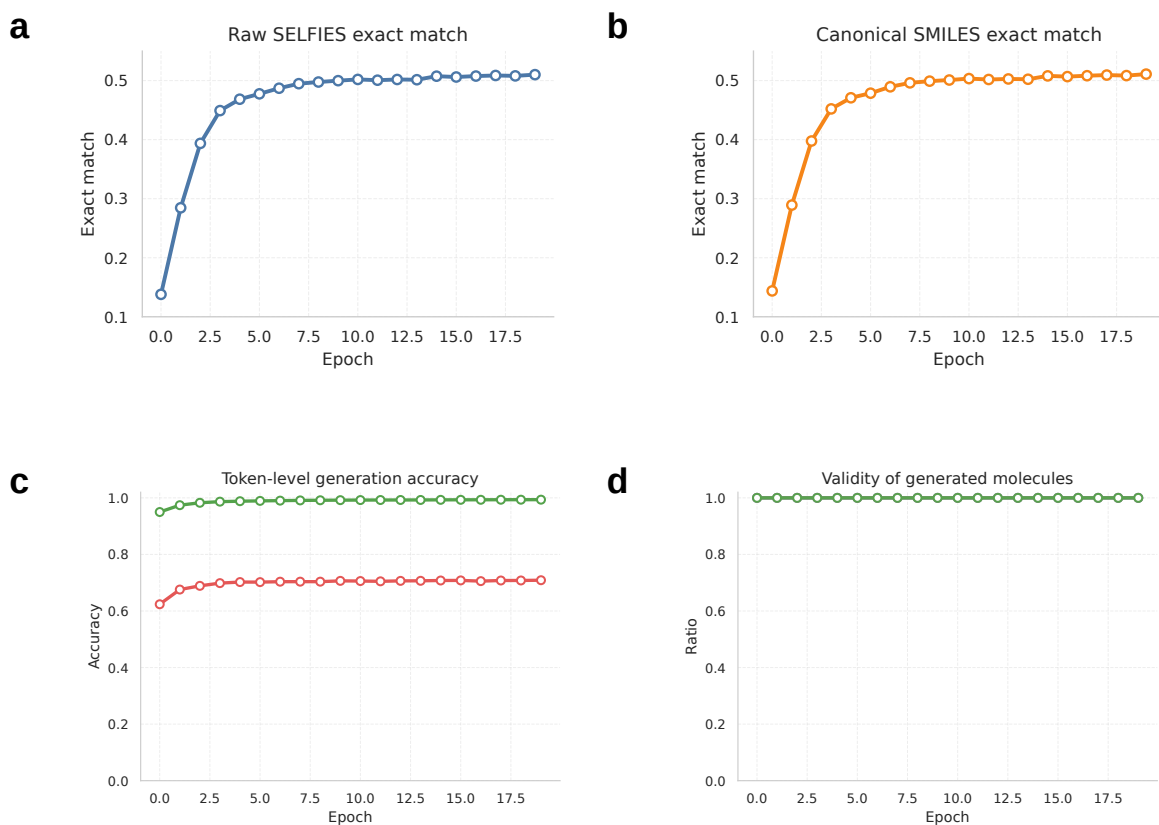


Figure 5: M2S music-to-SELFIES training and evaluation. The M2S route uses REMI/BPE music tokens¹⁸ as source sequences and shifted SELFIES token IDs as autoregressive decoder targets. Generated molecules are evaluated by raw SELFIES exact match, canonical SMILES exact match, token-level accuracy, and molecular validity. Across training, token-level generation accuracy increases to above 0.95, molecular validity approaches 1.0, and exact-match metrics improve steadily on held-out music-molecule pairs.

Music-derived representations preserve bioactivity neighborhoods. We next asked whether the neighborhood-aligned music representation captures chemically meaningful information beyond sonification. On 30 MOLECULEACE activity-prediction tasks, we used the music-derived piano-roll representation as a frozen feature extractor and fitted Ridge regression probes on the provided training splits. The representation was compared with four established molecular encoders spanning complementary modalities: SMILES-based ChemBERTa, graph-based GROVER, image-based ImageMol, and 3D-geometry-based Uni-Mol (Figure 6).

Across the 30 tasks, MUSICMOL achieved higher per-dataset Spearman correlation and R^2 , and lower RMSE, than all four baselines (Figure 6a). Paired two-sided t -test P -values ranged from 9.4×10^{-7} to 9.9×10^{-15} across metrics and baselines, indicating that the music-derived representation makes property-relevant molecular information accessible to simple linear probes.

To evaluate local neighborhood organization independently of downstream probe fitting, we computed the local label incoherence index (LII) for each assay dataset. LII measures the mean absolute difference in z-score-standardized assay values between each molecule and its k nearest neighbors in the frozen representation space; lower LII indicates that representation-space neighbors have more similar bioactivity. MUSICMOL maintained the lowest mean LII across all neighborhood sizes from $k=5$ to $k=100$ (Figure 6b). The advantage was most pronounced at smaller neighborhood sizes, the regime most relevant to local chemical neighborhoods and activity-cliff behavior. Per-dataset LII distributions at $k=5$, 10, and 20 further showed that the improvement was consistent across assay targets rather than being driven by a small subset of datasets (Figure 6c). Together, the linear-probe and LII analyses show that neighborhood-aligned music encodings retain both globally predictive and locally coherent molecular information.

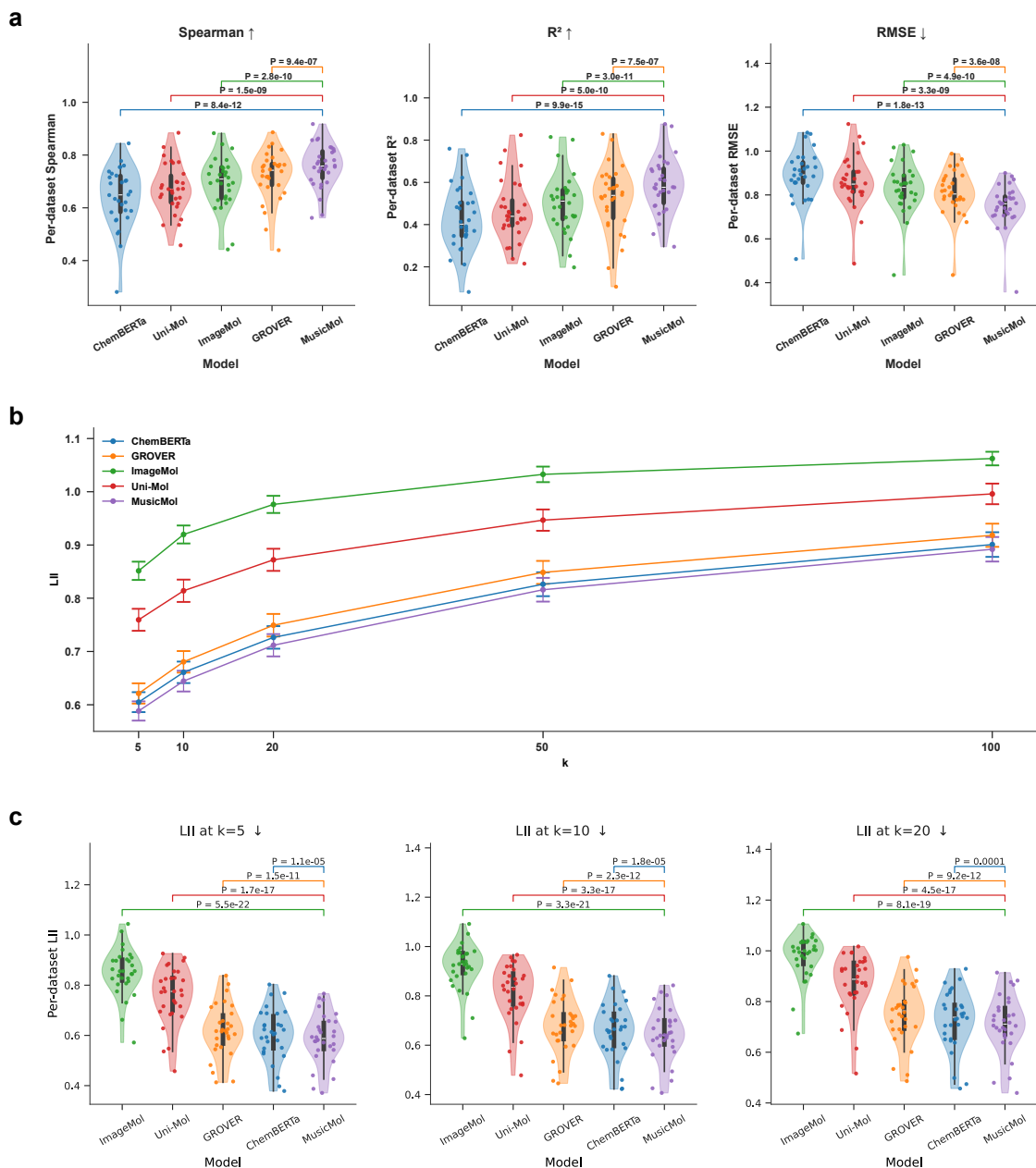


Figure 6: Benchmarking the music-derived molecular representation. **a**, Per-dataset frozen linear-probe performance on 30 MOLECULEACE assay tasks for ChemBERTa, Uni-Mol, ImageMol, GROVER, and MUSICMOL, summarized by Spearman correlation, R^2 , and RMSE. Higher Spearman correlation and R^2 , and lower RMSE, indicate better predictive performance; brackets show paired two-sided t -test P -values comparing MUSICMOL with each baseline over matched datasets. **b**, Mean local incoherence index (LII) on the MOLECULEACE collection as neighborhood size k varies from 5 to 100; error bars show the standard error across datasets. Lower LII indicates stronger preservation of local assay-value neighborhoods in representation space. **c**, Per-dataset LII distributions at $k=5$, 10, and 20. Violin plots show dataset-level values with internal box plots and individual dataset points; brackets indicate paired two-sided t -tests comparing MUSICMOL with each baseline ($n=30$ datasets per comparison).

Closed-loop interactive deployment of MUSICMOL. To examine whether the bidirectional MUSICMOL workflow could operate outside offline benchmarking, we deployed the system as an on-site interactive installation at the Future Learning Center of Zhejiang University ([Figure 7](#)). The installation presents the full molecule–music loop through a shared large-screen interface and an electronic keyboard. Users can select predefined molecular examples for immediate molecule-to-music conversion, edit or draw custom molecular structures with synchronized three-dimensional preview, and interact with the music-to-molecule module in which musical input is decoded into candidate molecular structures.

This deployment provides an operational test of the complete S2M–M2S pipeline in a public exhibition environment. The system must integrate molecular rendering, symbolic score generation, piano playback, real-time input handling, and model-based molecular generation within a single interactive workflow. It also demonstrates the educational potential of molecule–music translation. By allowing visitors to see, hear, and manipulate molecular representations through a familiar musical interface, the platform offers a perceptual entry point into molecular similarity and chemical space without requiring prior expertise in cheminformatics. Together, the S2M and M2S routes form a closed-loop interface in which molecules can be rendered as playable music and user-generated musical input can be translated into molecular candidates.

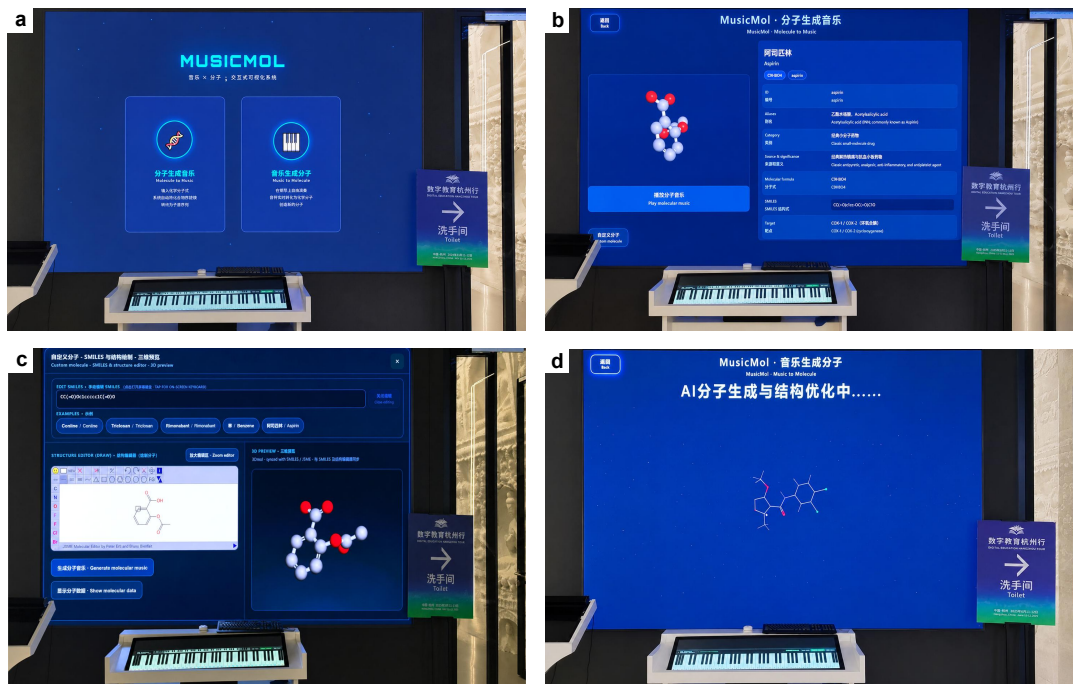


Figure 7: On-site interactive demonstration of the MUSICMOL platform at Zhejiang University. **a**, Physical deployment at the Future Learning Center, showing the shared large-screen interface and the electronic keyboard used for interactive input and playback. **b**, Molecule-to-music interface for a representative compound, displaying its three-dimensional molecular structure and molecular metadata before piano-based playback. **c**, Custom molecular input module supporting SMILES editing, two-dimensional structure drawing, and synchronized three-dimensional conformer preview prior to music generation. **d**, Music-to-molecule interface, in which musical input is processed to generate candidate molecular structures.

Discussion

MUSICMOL shows that symbolic music can function as a bidirectional interface for molecular representation and generation, rather than merely as a sonification output. By aligning molecular similarity neighborhoods with piano-roll music neighborhoods, the S2M route produces music representations that retain chemically meaningful local structure. This is supported by the MOLECULEACE linear-probe results and the lower local incoherence index across neighborhood sizes, indicating that nearby molecules in the learned music space tend to have more similar bioactivity profiles.

The reverse M2S route extends this interface from representation to generation. Using S2M-generated molecule-MIDI pairs, the Transformer decoder learns to translate symbolic music into valid SELFIES molecular strings, forming a closed loop in which molecules can be rendered as

music and music can be translated back into molecular candidates. The interactive deployment further demonstrates that this loop can operate with real-world musical inputs, including user-composed or performed melodies, making molecular generation accessible through a perceptual and creative modality.

Several limitations remain. The current representation evaluation is restricted to frozen encodings on MOLECULEACE, and broader benchmarking on additional property-prediction and generative-design tasks is needed. The M2S model is trained on paired data generated by S2M rather than naturally occurring molecule–music pairs, so its behavior on diverse external music corpora should be characterized more systematically. Moreover, although the generated SELF-IES strings are chemically valid, the current model does not yet explicitly optimize synthesizability, drug-likeness, novelty, diversity, or target-specific activity.

These limitations also define clear directions for music-guided molecular design. Future work could examine whether musical attributes such as tempo, rhythm density, pitch contour, harmony, style, emotional valence, or composer-specific patterns induce reproducible preferences in the generated molecular distributions. For example, music from different genres, emotional categories, or musicians could be mapped to molecular candidates and compared by scaffold composition, physicochemical properties, predicted bioactivity, and chemical novelty. On the design side, M2S can be coupled with target-specific activity predictors, docking scores, pharmacophore constraints, synthesizability filters, or multi-objective optimization to steer music-conditioned generation toward molecules for specific proteins, pathways, or phenotypic assays. In this sense, MUSICMOL provides a first representational and generative bridge between music and molecules, while target-aware music-guided molecular design remains an important next step.

Methods

The Methods describe MUSICMOL as a bidirectional framework with two linked but distinct routes. The structure-to-MIDI route (S2M) maps molecular structures to symbolic piano music through rule-guided conditional flow matching and neighborhood alignment, whereas the MIDI-to-structure route (M2S) maps symbolic music back to molecular sequences through supervised music-to-SELFIES decoding. We first describe dataset construction (Section 1), followed by S2M generation (Section 2), M2S decoding (Section 3), and representation benchmarking (Section 4).

1. Datasets and Data Construction

1.1. S2M training corpus

The S2M route was trained on two million compounds from ChEMBL. Each molecule was canonicalized with RDKit, standardized to a single connected molecular component when possible, and converted into a molecular graph for atom-level encoding. Invalid or empty molecules were excluded. S2M training did not use activity or property labels; instead, molecular graphs provided the input structures, CHEMMUSICRULE provided rule-guided warm-start note targets, and Morgan fingerprint distances provided the molecular neighborhood structure used in the alignment loss.

1.2. M2S molecule–MIDI paired corpus

Because large-scale experimental molecule–music paired data do not exist, the M2S corpus was constructed by applying the trained S2M route to selected PubChem molecules. Each molecule was first canonicalized and converted into a symbolic score using S2M. The resulting music sequence was then paired with the corresponding molecular SELFIES string. The music side provided source sequences for the M2S encoder, and the SELFIES side provided autoregressive target sequences for the decoder. This procedure yielded ten million compound–score pairs for supervised music-to-molecule learning.

1.3. MOLECULEACE benchmark

For downstream representation evaluation, we used the MOLECULEACE benchmark. The benchmark contains 30 assay tasks with 48,714 molecules in total, using the official train/test split for each task, including 38,912 training molecules and 9,802 test molecules in total (Table 1). For each validation seed (42, 43, and 44), 20% of the official training split was held out as a validation set, and the remaining molecules were used as the probe-training subset. Ridge regression probes were fitted on the probe-training subset, and the Ridge regularization strength was selected by validation MSE over 25 logarithmically spaced values from 10^{-3} to 10^6 .

Table 1: MOLECULEACE task sizes and provided splits.

Dataset	Total	Train	Test
CHEMBL1862_Ki	794	633	161
CHEMBL1871_Ki	659	525	134
CHEMBL2034_Ki	750	598	152
CHEMBL2047_EC50	631	503	128
CHEMBL204_Ki	2754	2201	553
CHEMBL2147_Ki	1456	1162	294
CHEMBL214_Ki	3317	2651	666
CHEMBL218_EC50	1031	823	208
CHEMBL219_Ki	1865	1491	374
CHEMBL228_Ki	1704	1362	342
CHEMBL231_Ki	973	776	197
CHEMBL233_Ki	3142	2512	630
CHEMBL234_Ki	3657	2924	733
CHEMBL235_EC50	2349	1877	472
CHEMBL236_Ki	2598	2077	521
CHEMBL237_EC50	955	762	193
CHEMBL237_Ki	2603	2081	522
CHEMBL238_Ki	1052	839	213
CHEMBL239_EC50	1721	1374	347
CHEMBL244_Ki	3097	2476	621
CHEMBL262_Ki	856	683	173
CHEMBL264_Ki	2862	2288	574
CHEMBL2835_Ki	615	489	126
CHEMBL287_Ki	1328	1061	267
CHEMBL2971_Ki	976	779	197
CHEMBL3979_EC50	1125	899	226
CHEMBL4005_Ki	960	767	193
CHEMBL4203_Ki	731	582	149
CHEMBL4616_EC50	682	543	139
CHEMBL4792_Ki	1471	1174	297

2. S2M: Molecule-to-Music Generation

For a molecule $G_k = (\mathcal{V}_k, \mathcal{E}_k)$, S2M produces atom-level note parameters and quantizes them into a symbolic score:

$$G_k \xrightarrow{\text{S2M}} \hat{\mathbf{Y}}_k = \{\hat{\mathbf{y}}_i\}_{v_i \in \mathcal{V}_k} \xrightarrow{\text{Quantize}} S_k, \quad \hat{\mathbf{y}}_i = (\hat{p}_i, \hat{s}_i, \hat{d}_i, \hat{v}_i) \in [0, 1]^4. \quad (1)$$

Here $\hat{p}_i$, $\hat{s}_i$, $\hat{d}_i$, and $\hat{v}_i$ denote normalized pitch, inter-onset step, duration, and velocity, respectively. In the differentiable piano-roll used for alignment, the step coordinate is treated as a normalized rhythmic coordinate. In the discrete symbolic score, the quantized step value advances the onset cursor, whereas duration controls note length after quantization.

2.1. Molecular graph encoding

Each SMILES string was canonicalized with RDKit and converted into a molecular graph². Atom features included element type, degree, hydrogen count, formal charge, hybridization, ring state, aromaticity, chirality, and local bond-type statistics. Bond features encoded bond type. A four-layer residual message-passing neural network^{4,5} mapped these features to atom-context vectors $\mathbf{h}_i \in \mathbb{R}^{256}$, which condition the S2M generator.

2.2. Rule-guided note targets

Each heavy atom was mapped to one note with four normalized parameters: pitch, inter-onset step, duration, and velocity. CHEMMUSICRULE provided warm-start targets for these parameters by using molecular topology and atom environments to form a musically coherent scaffold before learned flow matching and neighborhood alignment refined the mapping. Separating inter-onset step from duration allowed the generated score to express rests, articulation, and denser rhythmic patterns instead of forcing all temporal information into a single duration coordinate.

The deterministic rule encoder $\mathcal{E}_{\text{rule}}$ produced atom-level warm-start targets for the four note coordinates. Pitch was derived from the atom’s breadth-first-search depth from a central atom, producing an arch-shaped melodic contour with small ring-level pitch variation. The central atom was chosen deterministically: among ring atoms, the atom with the largest degree was used; if the molecule had no ring atom, the highest-degree atom in the molecular graph was used, with ties resolved by canonical atom order. Disconnected salts or mixtures were standardized to a single connected molecular component before applying the rule, and invalid or empty molecules were excluded.

Step, duration, and velocity were normalized functions of interpretable atom attributes. Step was determined mainly by ring size, aromaticity, terminal atoms, branch points, and bond order; aromatic rings received short regular steps, terminal atoms received longer steps, and branch points received shorter ornamental steps. Duration was defined as a step-scaled articulation value, with aromatic and terminal atoms assigned slightly shorter articulations than ordinary chain atoms. Velocity reflected structural importance, assigning larger values to ring junctions and branch points and smaller values to terminal atoms.

2.3. Conditional flow matching

S2M learns a conditional vector field from atom context using flow matching¹⁹. For each atom, noise $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_4)$ and time $t \sim \mathcal{U}(0, 1)$ define a straight-line path to the rule target:

$$\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\mathbf{y}_i^{\text{rule}}, \quad \mathbf{u}_t = \mathbf{y}_i^{\text{rule}} - \mathbf{z}_0. \quad (2)$$

The vector field is an MLP over the current point, a sinusoidal time embedding, and the atom context:

$$v_\theta(\mathbf{z}_t, t \mid \mathbf{h}_i) = \text{MLP}_\theta([\mathbf{z}_t \parallel \gamma(t) \parallel \mathbf{h}_i]). \quad (3)$$

The flow-matching loss is

$$\mathcal{L}_{\text{FM}} = \frac{1}{N} \sum_i \|v_\theta(\mathbf{z}_t, t \mid \mathbf{h}_i) - \mathbf{u}_t\|_2^2. \quad (4)$$

2.4. Neighborhood-alignment loss

The alignment loss follows the notation in [Figure 2](#). The target-space embeddings are L2-normalized ECFP4 fingerprints, and the predicted-space embeddings are L2-normalized soft piano-roll vectors. Generated atom parameters are rendered into a differentiable soft piano-roll to define the predicted music distance matrix D_{pred} . Let

$$\mu_i^p = \hat{p}_i(n_{\text{pitch}} - 1), \quad \mu_i^s = \hat{s}_i(n_{\text{time}} - 1). \quad (5)$$

The soft roll for molecule G_k is

$$\tilde{P}_k[p, \tau] = \sum_{v_i \in G_k} \hat{v}_i \exp\left(-\frac{(p - \mu_i^p)^2}{2\sigma_p^2}\right) \exp\left(-\frac{(\tau - \mu_i^s)^2}{2\sigma_\tau^2}\right), \quad (6)$$

and the music vector is

$$\mathbf{r}_k^{\text{roll}} = \frac{\text{vec}(\tilde{P}_k)}{\|\text{vec}(\tilde{P}_k)\|_2 + \epsilon}. \quad (7)$$

The music distance in a mini-batch is

$$D_{\text{pred},kl} = D_{kl}^{\text{music}} = 1 - \mathbf{r}_k^{\text{roll}\top} \mathbf{r}_l^{\text{roll}}. \quad (8)$$

In parallel, molecule-space distances were computed from RDKit Morgan fingerprints with radius 2 and 2048 bits, corresponding to an ECFP4-style representation^{20,21}. Each binary fingerprint was L2-normalized, and the off-diagonal molecule distance was the cosine distance:

$$D_{\text{target},kl} = D_{kl}^{\text{ECFP4}} = 1 - \frac{\mathbf{f}_k^\top \mathbf{f}_l}{\|\mathbf{f}_k\|_2 \|\mathbf{f}_l\|_2 + \epsilon}, \quad D_{\text{target},kk} = 0. \quad (9)$$

The target and predicted distance matrices were converted into row-stochastic neighborhood probabilities with a softmax kernel:

$$P_{\text{target},kl} = \frac{\exp(-D_{\text{target},kl}/\tau)}{\sum_{m \neq k} \exp(-D_{\text{target},km}/\tau)}, \quad P_{\text{pred},kl} = \frac{\exp(-D_{\text{pred},kl}/\tau)}{\sum_{m \neq k} \exp(-D_{\text{pred},km}/\tau)}. \quad (10)$$

The diagonal entries were set to zero, giving the figure quantities $P_{\text{target}} = K(D_{\text{target}})$ and $P_{\text{pred}} = K(D_{\text{pred}})$. The probability-alignment loss is a KL divergence from the fixed molecular neighborhood distribution to the generated-music neighborhood distribution:

$$\mathcal{L}_{\text{prob}} = \frac{1}{B} \sum_k \sum_{l \neq k} P_{\text{target},kl} \log \frac{P_{\text{target},kl}}{P_{\text{pred},kl} + \epsilon}. \quad (11)$$

A rank loss further encouraged the nearest ECFP4 neighbors of each molecule to remain closer than sampled distant molecules in music space:

$$\mathcal{L}_{\text{rank}} = \frac{1}{B} \sum_k \frac{1}{|\mathcal{P}_k| |\mathcal{N}_k|} \sum_{p \in \mathcal{P}_k} \sum_{n \in \mathcal{N}_k} \max(0, m + D_{\text{pred},kp} - D_{\text{pred},kn}), \quad (12)$$

where $\mathcal{P}_k$ contains the nearest molecular neighbors of anchor k under D_{target} and $\mathcal{N}_k$ contains sampled farther molecules. The figure terms d_{pos} and d_{neg} correspond to $D_{\text{pred},kp}$ and $D_{\text{pred},kn}$, respectively. The alignment loss is

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{prob}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}}. \quad (13)$$

2.5. S2M training objective and schedule

The S2M objective combines flow matching and neighborhood alignment:

$$\mathcal{L}_{\text{S2M}} = \lambda_{\text{FM}} \mathcal{L}_{\text{FM}} + \lambda_{\text{align}}(s) \mathcal{L}_{\text{align}}. \quad (14)$$

The implementation warms up the alignment term and decays the flow-matching supervision weight:

$$\begin{aligned} \lambda_{\text{align}}(s) &= \lambda_{\text{align}}^{\max} \min \left(1, \frac{s}{T_{\text{align}}} \right), \\ \lambda_{\text{FM}}(s) &= \lambda_{\text{FM}}^{\text{init}} + (\lambda_{\text{FM}}^{\text{final}} - \lambda_{\text{FM}}^{\text{init}}) \\ &\quad \times \min \left(1, \frac{\max(0, s - T_{\text{warm}})}{T_{\text{decay}}} \right). \end{aligned} \quad (15)$$

This schedule first lets the vector field learn the CHEMMUSICRULE target path, then shifts training pressure toward aligning molecular and music neighborhoods. No extra penalty is applied between generated notes and the rule targets beyond the flow-matching objective.

Before mini-batch optimization, molecular fingerprints and rule-guided note targets were precomputed. Each mini-batch was encoded with the GNN to obtain atom-context vectors. The conditional vector field was trained with flow-matching supervision, while generated note parameters were rendered into soft piano-rolls for the neighborhood-alignment loss. The GNN encoder and vector field were updated jointly using $\mathcal{L}_{\text{S2M}}$.

2.6. Score generation

At inference time, atom-level latent variables were initialized from a standard Gaussian distribution and transported by the learned vector field using a fixed-step explicit Euler solver. For evaluations, one sample was generated per molecule, and the vector field was integrated from $t = 0$ to $t = 1$ using 10 Euler steps. The resulting continuous atom-level parameters were clamped to $[0, 1]$ and quantized into a fixed pitch palette, inter-onset step values, duration buckets, and MIDI velocity.

The resulting notes were assembled into a `music21.Score` and exported as MIDI or MusicXML. The soft piano-roll uses $\hat{s}_i$ as a normalized rhythmic coordinate, whereas the discrete score uses the quantized step value to advance the onset cursor and the quantized duration value to set note length.

3. M2S: Music-to-Molecule Decoding

M2S performs the reverse translation. A symbolic music file, whether produced by S2M or supplied externally, is tokenized with a REMI/BPE music vocabulary, encoded by a sequence-to-sequence Transformer, and decoded autoregressively into SELFIES molecular tokens. The predicted SELFIES sequence is then converted into a molecular candidate. This design separates music generation and music-to-molecule decoding while keeping both routes connected through a shared symbolic music representation.

3.1. Music tokenization

M2S follows the route shown in [Figure 4](#). A symbolic music file, such as MIDI or MusicXML, was first converted into a REMI-style event sequence¹¹. The event sequence represents musical attributes such as bar position, note onset, pitch, duration, velocity, program, chord, tempo, and time signature as discrete tokens. Byte-pair encoding¹⁸ was then applied over these event tokens to obtain compact music token IDs. This tokenization makes externally supplied music and S2M-generated scores compatible with the same M2S encoder.

3.2. SELFIES target tokenization

The molecular target paired with each music sequence was represented as a SELFIES string rather than as a raw SMILES string, so that decoded sequences follow a chemically valid molecular grammar³. During training, the decoder input was the shifted SELFIES token sequence, implementing teacher forcing for autoregressive molecular-sequence learning. Padding masks were applied to ignore padded positions during loss computation.

3.3. Encoder–decoder Transformer

M2S uses an encoder–decoder Transformer to translate music-token sequences into SELFIES-token sequences¹⁷. Music token IDs are embedded and combined with sinusoidal positional encodings, then processed by the encoder to produce contextual music representations. The decoder attends to this encoder memory through cross-attention and predicts the next SELFIES token autoregressively using masked self-attention. The model is trained with token-level cross-entropy over shifted SELFIES targets.

Figure 5 summarizes the M2S training and evaluation route. Music token IDs are used as source sequences, shifted SELFIES token IDs provide autoregressive decoder targets, and generated molecules are evaluated by exact-match, token-level, and molecular-validity metrics.

3.4. Autoregressive SELFIES decoding

At inference time, M2S encodes an input musical sequence and generates SELFIES tokens autoregressively until an end-of-sequence token is produced or the maximum sequence length is reached. The predicted SELFIES token sequence is then converted into a canonical molecular string. This defines M2S as a music-to-molecule generation route: S2M-generated scores, user-composed melodies, or externally supplied symbolic music can be translated into predicted molecular candidates.

3.5. M2S evaluation metrics

M2S generation was evaluated on held-out molecule–music pairs. Raw SELFIES exact match was computed by comparing the generated SELFIES token sequence with the target SELFIES sequence. Canonical SMILES exact match was computed after decoding both generated and target SELFIES strings into molecular structures and canonicalizing them to SMILES. Token-level accuracy measured position-wise agreement between predicted and target SELFIES tokens over non-padding positions. Molecular validity was computed as the fraction of generated sequences that could be decoded into valid molecular structures.

4. Representation Benchmarking and Local Neighborhood Analysis

4.1. Frozen representations and baseline encoders

For MOLECULEACE evaluation, all models were used as frozen representation extractors using the datasets and splits described above. No baseline encoder or MUSICMOL encoder was fine-tuned during probe training, and no post-hoc calibration was applied. Test-set performance was reported using Spearman correlation, R^2 , and RMSE, then summarized across the three validation seeds.

The frozen representations compared in the linear probe were ChemBERTa, GROVER, ImageMol, Uni-Mol, and MUSICMOL (Table 2). The four baseline encoders were selected to cover representative molecular representation modalities. ChemBERTa represents SMILES pretraining with Transformer encoders ²²; GROVER represents graph pretraining with message-passing and Transformer-style molecular encoders ²³; ImageMol represents molecular image pretraining from rendered two-dimensional depictions ²⁴; and Uni-Mol represents three-dimensional molecular pretraining with geometry-aware Transformer representations ⁷. This panel compares the proposed music-derived representation against string, graph, image, and 3D-geometry baselines under the same downstream probe. Baseline implementations followed the corresponding public model releases and their recommended default pretrained checkpoints for ChemBERTa, GROVER, ImageMol, and Uni-Mol. For all representations, features were centered using statistics from the probe-training subset before Ridge fitting, without feature-wise variance scaling.

Table 2: Frozen representations used in MOLECULEACE linear probes.

Representation	Model	Feature extraction	Dimension
SMILES_based	ChemBERTa ²²	CLS pooling	768
graph_based	GROVER ²³	atom+bond fingerprint	4800
image_based	ImageMol ²⁴	ResNet18 encoder	512
3D_based	Uni-Mol ⁷	CLS representation	512
MIDI_based	MUSICMOL	soft piano-roll	5632

Checkpoints or versions:

ChemBERTa: seyonec/ChemBERTa-zinc-base-v1; GROVER: GROVER_large;

ImageMol: ImageMol.pth.tar; Uni-Mol: mol_pre_no_h_220816.pt; MUSICMOL: MusicMol.

4.2. Ridge linear probes

For each validation seed (42, 43, and 44), 20% of the provided training split was held out for validation, and the remaining training molecules were used to fit a Ridge regression probe. The Ridge regularization strength was selected by validation MSE over 25 logarithmically spaced values from 10^{-3} to 10^6 . The selected model was evaluated on the official test split for each dataset. Performance was reported using Spearman correlation, R^2 , and RMSE, and per-dataset values were used for paired statistical comparisons.

4.3. Local incoherence index

To evaluate whether each representation preserves local assay-value neighborhoods, we computed a local label incoherence index (LII) ²⁵ for each assay dataset. For molecule i , let $\mathbf{z}_i$ be the representation vector being evaluated and let y_i be the corresponding assay value. Assay values were standardized within each dataset before LII calculation:

$$\tilde{y}_i = \frac{y_i - \mu_y}{\sigma_y}. \quad (16)$$

For a neighborhood size k , $\mathcal{N}_k(i)$ denotes the k nearest neighbors of molecule i in representation space, excluding i itself. Neighbors were selected by cosine distance. The dataset-level LII is

$$\text{LII}_k = \frac{1}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_k(i)|} \sum_{j \in \mathcal{N}_k(i)} |\tilde{y}_j - \tilde{y}_i|. \quad (17)$$

Lower LII indicates that nearby molecules in representation space have more similar assay values. We computed LII for ChemBERTa, GROVER, ImageMol, Uni-Mol, and MUSICMOL on the same 30 MOLECULEACE assay datasets, using per-dataset z -score label standardization, cosine-neighbor retrieval, and $k \in \{5, 10, 20, 50, 100\}$. For MUSICMOL, representations were 88×64 soft piano-roll vectors (5,632 dimensions) extracted from the frozen S2M generator with ten Euler integration steps.

4.4. Statistical analysis

For the bracketed comparisons in [Figure 6](#), the unit of statistical analysis was the MOLECULEACE task. For each metric, MUSICMOL and a baseline were paired by dataset, and only matched dataset-level values were compared. In the linear-probe panel, the paired values were the per-dataset test-set Spearman correlation, R^2 , and RMSE values recomputed from saved test predictions. In the LII panel, the paired values were the per-dataset LII values at the displayed neighborhood size. Each comparison used a two-sided paired t -test computed in SciPy, with missing or

nonfinite values omitted; comparisons with fewer than three matched datasets were not annotated. Reported bracket labels are nominal P values without multiple-comparison correction. Violin plots show the distribution of dataset-level values, with internal box plots and overlaid individual dataset points.

Data and code availability. The data and code supporting this study are available at <https://github.com/shenwxlab/MusicMol/tree/musicmol-v1>. The full source code will be made available upon publication. The raw ChEMBL data are available from <https://www.ebi.ac.uk/chembl/>. The raw PubChem compound data are available from <https://pubchem.ncbi.nlm.nih.gov/docs/downloads> and <https://ftp.ncbi.nlm.nih.gov/pubchem/>. The MoleculeACE benchmark is available from <https://github.com/molML/MoleculeACE>.

Acknowledgements. This work was supported by the “ZJU100 Young Professor” program of Zhejiang University, China. This work was carried out with substantial assistance in computation and infrastructure, and was supported by Information Technology Center, Zhejiang University, and China Mobile Zhejiang Co., Ltd. Hangzhou Branch.

Author contributions. Conceptualization, W.S., J.T., and C.C.; methodology, W.S., J.T., Z.Z., and C.C.; software, J.T., C.C., N.K., P.J., C.Z., and Z.Y.; investigation, J.T., Z.Z., C.C., N.K., C.S., Y.L., Y.Z., and C.Y.; data curation, J.T., Z.Z., N.K., C.S., Y.L., Y.Z., and C.Y.; formal analysis, J.T., Z.Z., C.C., N.K., and C.S.; visualization, J.T., Z.Z., C.C., and N.K.; writing—original draft, J.T., Z.Z., C.C., and W.S.; writing—review and editing, all authors; resources, W.S.; supervision, W.S.; funding acquisition, W.S.

Competing interests. The authors declare that they have no known competing financial interests.

References

1. Nguyen-Vo, T.-H., Teesdale-Spittle, P., Harvey, J. E. & Nguyen, B. P. Molecular representations in bio-cheminformatics. *Memetic Computing* **16**, 519–536 (2024).
2. Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28**, 31–36 (1988).
3. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **1**, 045024 (2020).
4. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, 1263–1272 (PMLR, 2017).
5. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations* (2019).
6. Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O. & Baker, N. Chemception: A deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv preprint arXiv:1706.06689* (2017). 1706.06689.
7. Zhou, G. *et al.* Uni-mol: A universal 3d molecular representation learning framework. In *International Conference on Learning Representations* (2023).
8. Hermann, T., Hunt, A. & Neuhoff, J. G. (eds.) *The Sonification Handbook* (Logos Verlag Berlin, 2011).
9. Braun, R., Tfirm, M. & Ford, R. M. Listening to life: Sonification for enhancing discovery in biological research. *Biotechnology and Bioengineering* **121**, 3009–3019 (2024).
10. Huang, C.-Z. A. *et al.* Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations* (2019).
11. Huang, Y.-S. & Yang, Y.-H. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1180–1188 (Association for Computing Machinery, 2020).
12. Zeng, M. *et al.* MusicBERT: Symbolic music understanding with large-scale pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 791–800 (Association for Computational Linguistics, Online, 2021).
13. Chou, Y.-H., Chen, I.-C., Ching, J., Chang, C.-J. & Yang, Y.-H. MidiBERT-Piano: Large-scale pre-training for symbolic music classification tasks. *Journal of Creative Music Systems* **8** (2024).
14. Kumbar, M. Musical chemistry: Integrating chemistry and music. a nine-part series on generating music from chemical processes. *Journal of Chemical Education* **84**, 1933–1936 (2007).
15. Yu, C.-H., Qin, Z., Martin-Martinez, F. J. & Buehler, M. J. A self-consistent sonification method to translate amino acid sequences into musical compositions and application in protein design using artificial intelligence. *ACS Nano* **13**, 7471–7482 (2019).

16. Mahjour, B., Bench, J., Zhang, R., Frazier, J. & Cernak, T. Molecular sonification for molecule to music information transfer. *Digital Discovery* **2**, 520–530 (2023).
17. Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).
18. Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with sub-word units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725 (Association for Computational Linguistics, Berlin, Germany, 2016).
19. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. In *International Conference on Learning Representations* (2023). 2210.02747.
20. Morgan, H. L. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *Journal of Chemical Documentation* **5**, 107–113 (1965).
21. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling* **50**, 742–754 (2010).
22. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885* (2020). 2010.09885.
23. Rong, Y. *et al.* Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, vol. 33, 12559–12571 (2020).
24. Zeng, X. *et al.* Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence* **4**, 1004–1016 (2022).
25. Shen, W. X. *et al.* Activity cliff-informed contrastive learning for molecular property prediction. *Research Square* rs-3 (2024).