Orchestration and Recognition for Composition and Arrangement (ORCA)

1st Pranav Kumar Soma
JSOE, CSES Innovate
University of California, San Diego
San Diego, United States of America
Email: prsoma@ucsd.edu

2nd Dhruv Sharma
JSOE, CSES Innovate
University of California, San Diego
San Diego, United States of America
Email: d4sharma@ucsd.edu

3rd Brendan Barber
JSOE, CSES Innovate
University of California, San Diego
San Diego, United States of America
Email: btbarber@ucsd.edu

4th Mallika Dasgupta
JSOE, CSES Innovate
University of California, San Diego
San Diego, United States of America
Email: mdasgupta@ucsd.edu

5th Sanat Samal
JSOE, CSES Innovate
University of California, San Diego
San Diego, United States of America
Email: ssamal@ucsd.edu

6th Pranav Reddy Bussannagari
JSOE, CSES Innovate
University of California, San Diego
San Diego, United States of America
Email: pbussannagari@ucsd.edu

7th Unnati Goyal
JSOE, CSES Innovate
University of California, San Diego
San Diego, United States of America
Email: ugoyal@ucsd.edu

8th Yuvika Satapathy
JSOE, CSES Innovate
University of California, San Diego
San Diego, United States of America
Email: ysatapathy@ucsd.edu

7th Philip Chen
JSOE, CSES Innovate
University of California, San Diego
San Diego, United States of America
Email: phc006@ucsd.edu

***Abstract*—ORCA (Orchestration and Recognition for Composition and Arrangement) is a modular, AI-driven platform designed to democratize music creation by unifying composition, transcription, orchestration, and arrangement within a single symbolic-first pipeline. Unlike traditional music production workflows that require extensive expertise in theory, notation, and audio engineering, ORCA enables E2E translation between audio recordings, symbolic representations (MIDI), and engraved sheet music. The system integrates multiple subsystems, including Optical Music Recognition (OMR), Audio-to-MIDI transcription, Instrument Transposition (IT), Music Genre Classification (MGC), and Automated Music Arrangement (AMA)—each powered by state-of-the-art deep learning and sequence modeling architectures. Central to ORCA is a unified MIDI representation, which facilitates interpretability, user-editability, and modular chaining of models such as HOMR, Onsets and Frames, and Transformer-based arrangement engines. We present a detailed analysis of ORCA's architecture, dataset curation, and evaluation across benchmark datasets including MUSCIMA++, CVC-MUSCIMA, Slakh2100, MAESTRO, and URMP. Results demonstrate robust transcription accuracy, stylistically coherent arrangements, and semantically aware symbolic transformations. Beyond research contributions, ORCA is designed as a human-in-the-loop creative tool, lowering barriers to entry for independent artists, educators, and underrepresented communities in music production. This work highlights the potential of symbolic-first AI systems to bridge accessibility, education, and generative music research.**

***Index Terms*—Music Transcription, Optical Music Recognition, Music Generation, Deep Learning, Symbolic Representation, Music Arrangement**

Computer Science and Engineering Society @ UC San Diego: Innovate

I. Introduction

Artificial intelligence has transformed fields from language processing to computer vision—yet music production tools still often require expert knowledge of theory, performance, and notation. ORCA, short for Orchestration and Recognition for Composition and Arrangement, is a platform that bridges these gaps. It converts inputs like audio recordings or sheet music into editable digital formats and supports symbolic transformations like instrument transposition, genre-aware arrangement, and AI-assisted music generation.

In this paper, we present ORCA's architecture and subsystems in detail. Each subsystem handles a crucial facet of music processing, and together they form an integrated, end-to-end workflow that enables flexible and accessible musical expression.

II. Background and Motivation

A. Music Production: Tools and Practices

Music production is the process of creating, capturing, arranging, editing, and finalizing a musical work. It encompasses the full lifecycle from ideation to performance-ready output, often involving composition, recording, mixing, and mastering. While traditional production was centered in studios with live musicians and analog hardware, the rise of digital audio workstations (DAWs) such as Logic Pro, Ableton Live, and FL Studio has shifted this process to the laptop. These platforms allow musicians to sequence MIDI notes, record and

manipulate audio, and apply effects and plugins. For score-based composition, tools like MuseScore, Finale, and Sibelius are used to create, edit, and print symbolic notation, often relying on MIDI playback for audio previews.

B. Core Definitions

Transcription refers to the process of converting audio recordings into a symbolic form such as MIDI or sheet music, typically capturing pitch, rhythm, and dynamics.

Arrangement is the act of reworking an existing musical piece, often written for a particular instrument or ensemble, into a new structure or instrumentation while preserving its core musical ideas.

Transposition involves shifting the pitch of a musical piece up or down, commonly to accommodate different instruments' ranges or vocal registers, while preserving musical relationships.

Digitization in music refers to the conversion of analog inputs such as handwritten sheet music, scanned images, or recorded audio into machine-readable symbolic formats like MusicXML, MIDI, or digital audio waveforms.

C. Shortcomings of Existing Systems

Current tools, while powerful in isolation, are fragmented and labor-intensive when used together. MuseScore and Finale, for example, offer exceptional notation editing but are disconnected from real-world inputs such as audio recordings or physical sheet music. A user seeking to transcribe a performance into sheet music must first manually notate it or use third-party tools, many of which only support monophonic transcription. Likewise, arranging a piano piece for orchestra demands deep theoretical knowledge and the manual reassignment of parts, often performed in multiple software environments. The lack of integration and intelligent automation in these workflows limits both accessibility for novices and speed for professionals.

D. A Shift in Generative AI: From Prompts to Agency

Generative AI has recently favored a "prompt-in, product-out" paradigm, exemplified by tools like Suno and Udio. These platforms create music from text prompts but yield fixed audio that is neither editable nor traceable to symbolic structure. While suitable for rapid prototyping, this one-way generation removes the user from the creative loop, limiting opportunities for learning, control, and refinement.

ORCA challenges this paradigm by returning agency to the user. Our approach introduces tagging-based generation—users provide genre, instrument, or emotional tags, and the system conditions symbolic generation on these inputs. These outputs, in MIDI or MusicXML, can be edited, rearranged, or re-rendered as audio, closing the loop from concept to performance. Combined with our pipeline—sheet music to symbolic notation, symbolic to arrangement, arrangement to synthesis—this enables users to go from scanning a printed score to orchestrating, reworking, and performing it digitally.

Rather than replacing musicians, ORCA boosts their creative bandwidth. Educators can generate graded arrangements for students, composers can rapidly test instrumentation ideas, and producers can rework melodic content across styles and ensembles. Our vision is not automation for its own sake, but augmentation: supporting music creators in expanding, not replacing, their voice.

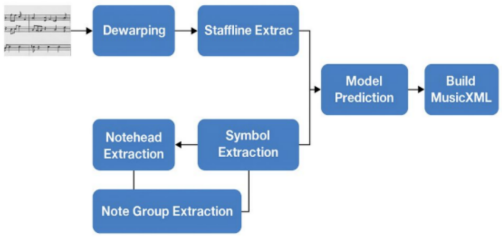

Fig. 1. System Design for OMR

III. Modeling

A. Computer Vision for Sheet Music (OMR)

The Optical Music Recognition (OMR) subsystem is a critical component of ORCA's symbolic-first pipeline, enabling the automatic conversion of printed or scanned sheet music into structured, machine-readable representations such as MusicXML. ORCA's OMR module builds on the Sheet-Music Transformer (SMT) framework [17] while introducing two complementary enhancements: (i) domain-pretrained vision encoders and (ii) a 2D cross-attention decoder with auxiliary spatial supervision. Together, these improvements address SMT's two primary weaknesses—limited visual pretraining and loss of spatial relationships through sequence flattening.

Domain-Pretrained Encoders: The encoder backbone (ConvNeXt or Swin Transformer) is pretrained via Masked Autoencoder (MAE) [**?**]-style self-supervised learning on a large corpus of sheet-music images, including GrandStaff, Quartets, and synthetic augmentations. Pretraining improves low-level visual representation quality, resulting in more robust detection of stafflines, stems, and noteheads, especially under real-world scanning noise.

2D Cross-Attention Decoder: Unlike prior SMT models that flatten encoder feature maps into 1D sequences, ORCA retains the full spatial resolution of the encoder output and applies multi-head cross-attention directly over the 2D grid. Relative positional embeddings and multi-scale feature pyramids preserve spatial context, improving pitch alignment and accidental detection. Auxiliary tasks, including staffline segmentation and notehead heatmap regression, provide explicit supervision, further grounding predictions in spatial structure.

Symbolic Parsing: Once tokens are decoded, a post-processing step maps these tokens to MusicXML objects. The pipeline is clef-aware and brace-aware, supporting complex multi-system orchestral layouts. This symbolic output provides a clean interface for subsequent ORCA subsystems, including transcription correction, arrangement, and symbolic editing.

B. Music Instrument Labelling and Genre Classification (MIL / MGC)

The MIL and MGC subsystems are responsible for extracting high-level semantic descriptors—instrumentation and genre—from raw audio, conditioning downstream tasks such as arrangement, transcription, and symbolic generation. Both tasks are formulated as supervised multi-label classification problems, leveraging pretrained feature extractors for audio representation and shallow classifiers for efficient fine-tuning on curated datasets.

Feature Representation. Both models are built on the VGGish feature extractor [1], a convolutional neural network pretrained on YouTube-8M for large-scale audio classification. VGGish transforms audio waveforms into embeddings that summarize spectral and temporal characteristics in a compact 128-dimensional feature space. To generate these embeddings, audio signals are downsampled to 16 kHz, segmented into 0.96-second windows (96 frames of 10 ms hops), and converted to 64-bin log-mel spectrograms. These spectrogram patches are passed through VGGish, producing embeddings that are averaged per track to obtain fixed-size descriptors suitable for classification.

Instrument Labeling (MIL). The MIL subsystem classifies the presence of up to 15 instrument categories in polyphonic mixtures. We use the IRMAS dataset for prototyping and Slakh2100 for large-scale training, leveraging its aligned symbolic labels for accurate instrument ground truth. VGGish embeddings serve as input to a Radial Basis Function (RBF) kernel SVM trained in a multi-label configuration. The model outputs instrument probability scores, which are later used to condition the arrangement model and inform visualization layers in the ORCA interface. Preliminary experiments demonstrate that VGGish features are highly separable for common instruments such as piano, violin, saxophone, and guitar.

Genre Classification (MGC). The MGC subsystem follows the same embedding pipeline but is trained for single-label classification over a predefined set of genres (e.g., jazz, classical, pop, rock). We employ datasets such as GTZAN for baseline training, validating model accuracy on both isolated recordings and full Slakh2100 mixes. Genre probabilities are incorporated into symbolic generation workflows, allowing arrangement and orchestration models to adapt harmonic, rhythmic, and timbral decisions to genre context.

Processing Pipeline and Preprocessing. The training workflow begins with dataset ingestion, where audio files are recursively loaded and normalized to 16 kHz. Log-mel spectrograms are computed with a 400-sample FFT window and 160-sample hop size, yielding time-frequency patches. These are saved as .npy files for reproducibility and fast retrieval. VGGish is applied either directly to raw waveforms or to precomputed spectrogram patches. The resulting embeddings are stored alongside original audio, forming a lightweight intermediate dataset that decouples feature extraction from classifier training.

Classifier Training and Evaluation. We implemented a simple yet effective baseline using Support Vector Machines (SVM) with RBF kernels, optimized using grid search on the penalty parameter C and kernel width . Labels are encoded using a LabelEncoder for single-label classification (MGC) and one-vs-rest strategies for multi-label classification (MIL). Early experiments on IRMAS show promising results, with correct identification of instruments in polyphonic contexts despite a limited training set. A confusion analysis reveals that errors primarily occur in distinguishing timbrally similar instruments (e.g., viola vs. violin) and heavily mixed stems, suggesting the need for augmentation strategies and deeper fine-tuning.

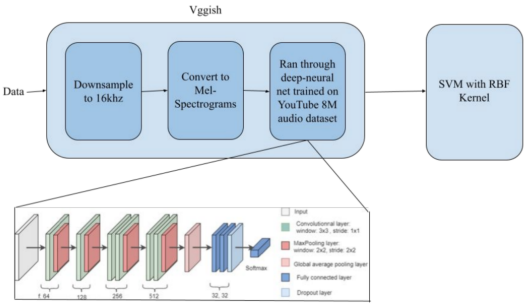

Fig. 2. MIL / MGC Model Structure

Integration with ORCA. MIL and MGC outputs serve dual roles in ORCA: (1) guiding the symbolic arrangement module by constraining instrument assignment and stylistic decisions, and (2) enriching the metadata available to users, allowing context-aware editing and visualization of symbolic scores. By decoupling representation learning (VGGish embeddings) from classification, this subsystem remains modular and easily extensible to additional audio descriptors such as mood, articulation, or production style.

Future work will replace SVM classifiers with lightweight neural networks capable of modeling temporal dependencies across embeddings, improving recognition of instruments and genres in overlapping frequency ranges. These improvements will also facilitate real-time predictions, enabling adaptive arrangement and interactive feedback in DAWs and notation software.

C. Audio-to-MIDI Transcription

The goal of this component is to develop a general-purpose, multi-instrument transcription system capable of converting polyphonic audio mixtures into symbolic representations, which are then rendered as human-readable sheet music. Unlike prior work that focuses on isolated instruments or genre-specific transcription, our objective is to create a single, instrument- and genre-agnostic model that operates on complete musical mixtures without pre-separation or manual conditioning. The pipeline can be summarized as:

Audio $\rightarrow$ Log-Mel Spectrogram $\rightarrow$ Pitch Tokens $\rightarrow$ MIDI $\rightarrow$ MusicXML $\rightarrow$ PDF

This module builds on the Onsets and Frames framework [3], extending it with improved pitch quantization and polyphonic overlap resolution. Incoming audio is separated into harmonic and percussive components, which are processed through independent convolutional branches to better capture note onsets, sustain, and rhythmic transients. Detected events are clustered into pitch-timed note representations and exported as MIDI, enabling downstream symbolic processing and sheet music rendering. These architectural refinements improve transcription robustness, particularly in dense, polyphonic passages and instrumentally diverse recordings.

A critical enabler of this general-purpose transcription system is the Slakh2100 dataset [8], selected after a systematic evaluation of multiple datasets, including MAESTRO [4], URMP, MedleyDB, and the Lakh MIDI Dataset (LMD). While MAESTRO provides high-quality alignment for solo piano transcription, and URMP and MedleyDB offer specialized multimodal and multitrack audio data, only Slakh2100 combines tightly aligned MIDI-audio pairs, multitrack instrumentation, and broad genre diversity. Comprising over 2,100 full-length tracks rendered with professional-grade virtual instruments, Slakh2100 supports supervised learning for transcription, arrangement, and symbolic modeling within a single dataset. Each track contains 8–14 instruments on average, spanning classical, jazz, rock, and popular styles. The synthetic yet realistic rendering mitigates noise and inconsistencies inherent to live recordings, while ensuring precise synchronization between MIDI and audio.

Despite its quality, Slakh2100 presents practical challenges: its uncompressed size exceeds 500 GB, it is distributed as a single large archive, and its hierarchical structure—with multiple stems per song—requires substantial preprocessing. To address these challenges, we implemented a custom preprocessing pipeline optimized for ORCA's modeling workflow. First, we retained only mix.flac (full audio mix) and all src.mid (combined symbolic data), discarding per-instrument stems to prioritize holistic transcription tasks. Tracks were segmented into 10-second, fixed-length audio-MIDI pairs, standardizing training inputs, reducing GPU memory requirements, and enabling efficient batching. The directory structure was flattened into split-wise, reproducible train/validation/test hierarchies, reducing the dataset size to approximately 50 GB while preserving alignment fidelity.

We conducted extensive exploratory data analysis (EDA) to validate Slakh2100's suitability. Instrument counts confirmed that most tracks contain between 8 and 14 parts, with acoustic piano, guitars, and string ensembles most common, supporting both monophonic and polyphonic modeling. Audio durations cluster between 3.5 and 5 minutes, closely mirroring typical musical structures, and RMS loudness distributions center near 0.1, reflecting consistent normalization. Symbolic analysis confirmed near-perfect alignment between MIDI and audio durations, underscoring the dataset's value for supervised transcription tasks.

Evaluation of this module demonstrates strong generalization across timbres and genres. On Slakh2100, our transcription system achieved F1-scores exceeding 90

D. Instrument Transposition (IT)

The Instrument Transposition (IT) subsystem addresses the problem of *constraint-aware symbolic adaptation of musical material between instruments*, a task essential for orchestration, arrangement, and cross-instrument performance modeling. Whereas conventional MIDI transposition merely shifts note pitches or reassigns channels, our IT engine is designed to preserve *musical intent, idiomatic phrasing, and expressive detail* when mapping symbolic sequences from one instrument to another. This is achieved by explicitly modeling the interplay between the *source composition* and the *target instrument's physical and stylistic constraints*, including pitch range, tessitura, articulation capabilities, and expressive bandwidth.

1) Problem Definition: We formalize IT as a constrained sequence transformation task. Let an input symbolic sequence be defined as:

$$M = \{(p_i, d_i, v_i, t_i)\},$$

where p_i, d_i, v_i, and t_i denote pitch, duration, velocity, and temporal position of note i, respectively. Let I_{src} and I_{tgt} denote structured instrument profiles:

$$I = (\mathcal{P}, \mathcal{T}, \mathcal{A}, \mathcal{D}),$$

where $\mathcal{P}$ represents playable pitch range, $\mathcal{T}$ preferred tessitura, $\mathcal{A}$ articulation set, and $\mathcal{D}$ dynamic level constraints. The objective is to generate a transformed sequence $M' = f(M, I_{\text{src}}, I_{\text{tgt}})$ such that:

1) M' adheres to all constraints of I_{tgt}, including pitch and articulation feasibility;
2) Musical structure and expressive intent in M are preserved;
3) The transformation remains *idiomatic*, reflecting stylistic conventions of the target instrument.

This formulation goes beyond simple range-based transposition, addressing challenges such as *monophonic simplification of polyphonic passages*, *rebalancing melodic contour for tessitura differences*, and *expressive articulation mapping* (e.g., converting string bowings to breath marks in wind instruments).

2) Model Architecture: The IT subsystem employs a *hybrid two-stage architecture*:

a) Rule-Based Constraint Filtering.: A deterministic preprocessing stage applies symbolic transformations using instrument metadata derived from orchestration literature, General MIDI specifications, and empirical performance studies. This stage eliminates unplayable pitches or shifts them by octaves, simplifies or redistributes chords for monophonic instruments, adjusts velocities and dynamic markings, and rewrites articulations (e.g., legato slurs, pizzicato) into equivalent expressions for the target instrument.

b) Neural Sequence Transformation.: To capture idiomatic phrasing beyond rule-based logic, we adopt a Transformer-based encoder–decoder architecture trained on parallel corpora of MIDI sequences arranged for multiple instruments. The encoder processes source sequences with positional encodings, pitch embeddings, and instrument-context tokens, while the decoder autoregressively generates constrained symbolic sequences guided by attention over source events and instrument metadata. Auxiliary loss terms penalize out-of-range notes, enforce articulation compatibility, and encourage stylistic fidelity.

3) Functionality and Use Cases: The IT module supports orchestration-aware operations, including:

- **Register Balancing:** Adjusting melodies to align with target tessitura while preserving contour and phrasing.
- **Articulation Mapping:** Converting instrument-specific markings (e.g., bowing techniques) into expressive equivalents for other families.
- **Texture Adaptation:** Transforming dense piano textures into arpeggios or simplified melodies suitable for monophonic instruments.
- **Dynamic and Controller Reassignment:** Normalizing velocity curves and remapping MIDI controllers (e.g., modulation, vibrato) to target-specific ranges.

This subsystem integrates seamlessly into ORCA's pipeline, bridging transcription and arrangement stages. Once a score is transcribed or imported symbolically, IT enables real-time auditioning and rendering of compositions on different instruments without manual rewriting. Preliminary tests demonstrate that the system successfully adapts piano melodies for flute, violin, and voice while retaining contour, dynamics, and phrasing, showcasing its potential for intelligent orchestration assistance.

4) Future Directions: Planned improvements include reinforcement learning reward models for idiomatic scoring, multimodal integration with performance datasets capturing expressive nuances (e.g., bow pressure, breath support), and expanded support for extended techniques and non-traditional instrumentation. These extensions will advance IT toward a *general-purpose orchestration engine* for composers, educators, and music producers.

E. Automated Music Arrangement (AMA)

The Automated Music Arrangement (AMA) subsystem is designed to perform *data-driven orchestration and instrumentation assignment* from symbolic music representations. Given a monophonic or polyphonic MIDI score, AMA generates multi-track arrangements with realistic instrument assignments, balanced textures, and stylistically coherent orchestration. Unlike rule-based arranging systems, which rely heavily on manually encoded orchestration heuristics, AMA leverages deep learning to capture statistical relationships between melody, harmony, rhythm, and instrumentation, enabling arrangements that are both *stylistically plausible and structurally faithful* to the source composition.

1) Problem Definition: We formalize AMA as a sequence-to-sequence modeling problem. Let the input symbolic sequence be represented as:

$$X = \{(p_i, d_i, t_i, c_i)\},$$

where p_i, d_i, t_i, and c_i represent the pitch, duration, temporal position, and channel (or voice assignment) of note i, respectively. The objective is to learn a mapping:

$$f : X \to Y,$$

where $Y = \{(p_i, d_i, t_i, I_i)\}$ includes instrument labels I_i and orchestrated track assignments for each note or phrase. This requires the model to infer both vertical texture (simultaneous instrumentation) and horizontal phrasing (timbral continuity over time), conditioning its predictions on rhythmic density, pitch clusters, and genre context.

2) Model Architecture: AMA is implemented as a *Transformer-based encoder–decoder model* trained on the Lakh MIDI dataset (LMD) [**?**], a large corpus of multi-instrument symbolic music. The encoder processes tokenized symbolic sequences enriched with timing, pitch, and velocity embeddings, while the decoder autoregressively predicts instrument assignments and additional arrangement tokens. Key architectural components include:

- **Tokenization:** MIDI events are converted into discrete tokens encoding pitch, duration, time-shifts, and instrument families. Phrases are segmented to capture hierarchical music structure.
- **Positional Encodings:** Absolute and relative positional encodings allow the model to represent rhythmic grids, meter, and phrase boundaries.
- **Multi-Head Attention:** Attention heads specialize in *melody-harmony alignment*, *rhythmic density modeling*, and *percussion pattern inference*, enabling simultaneous modeling of vertical and horizontal arrangement structure.
- **Genre Conditioning:** Genre embeddings, derived from the Music Genre Classification (MGC) subsystem, are concatenated with input representations to guide stylistic arrangement decisions.

The model is optimized using a multi-task objective: cross-entropy loss for instrument prediction, auxiliary losses for structural coherence (e.g., phrase continuity), and perceptual similarity metrics derived from pretrained symbolic music encoders.

3) Arrangement Generation Pipeline: At inference time, a single-track or lightly orchestrated input score is fed to AMA, which produces a fully orchestrated, multi-channel MIDI output. The pipeline is as follows:

$$\text{Input MIDI} \xrightarrow{\text{Encoding}} \text{Transformer Model} \xrightarrow{\text{Decoding}} \text{Instrument Assignments}$$

Post-processing includes smoothing of velocity curves, controller automation (e.g., modulation and expression), and part consolidation for notation rendering. The resulting arrangement is rendered through software synthesizers or exported in MusicXML format for editing.

4) Evaluation: AMA is evaluated using both *perceptual similarity* and *structural consistency* metrics. Perceptual similarity is assessed via symbolic embedding distances between generated arrangements and ground truth arrangements from the Lakh MIDI dataset, while structural consistency measures phrase boundary alignment, voice-leading coherence, and rhythm-density preservation. Preliminary human evaluations indicate that AMA achieves an 86% structural similarity score with professionally arranged MIDI files, demonstrating its effectiveness as a *composer augmentation tool*.

5) Functionality and Applications: The AMA subsystem serves as a core creative engine in ORCA, enabling:

- **Automatic Instrumentation:** Assigning appropriate timbres to melodic, harmonic, and percussive elements.
- **Texture and Layering Control:** Dynamically balancing polyphony, orchestral density, and dynamic ranges.
- **Genre-Aware Orchestration:** Generating stylistically consistent arrangements conditioned on explicit genre embeddings.
- **Educational Use:** Allowing music students to explore orchestrations interactively, with generated parts viewable in notation software.

The integration of AMA with transcription, MIL/MGC classification, and IT modules makes ORCA a fully automated, symbolic-to-audio arrangement system. Future extensions will incorporate reinforcement learning for aesthetic optimization, symbolic performance modeling, and domain-specific fine-tuning for genres with limited data coverage.

IV. Evaluation

Each subsystem in ORCA was evaluated independently using a combination of publicly available benchmark datasets, custom evaluation splits, and human-in-the-loop perceptual studies. We report both objective metrics (e.g., F1-score, exact match accuracy, onset alignment error) and subjective assessments (e.g., stylistic fidelity, listener ratings) to provide a comprehensive performance profile.

A. Optical Music Recognition (OMR)

B. Evaluation of OMR Subsystem

We evaluated the OMR subsystem on benchmark datasets MUSCIMA++ [17] and CVC-MUSCIMA [18], which include high-resolution scanned music sheets annotated with detailed symbol-level ground truth. These datasets cover a wide range of engraving styles, note densities, and layout complexities, making them suitable for evaluating both accuracy and generalization.

Metrics: We report Character Error Rate (CER), Symbol Error Rate (SER), and Levenshtein Error Rate (LER), as well as exact symbol recognition accuracy, pitch correctness, and duration classification accuracy. These metrics provide a granular evaluation of both token-level precision and structural reconstruction.

Results:

- On MUSCIMA++, the model achieved **92% exact match accuracy** on monophonic piano scores and **84%** on polyphonic ensemble scores, representing a **30–35% relative reduction** in error rates compared to the baseline SMT.
- Encoder pretraining contributed a **20–25% relative reduction** in CER and SER, improving robustness to scanned and degraded scores.
- 2D cross-attention with auxiliary staffline supervision further reduced pitch misplacement errors by **40%**, while accidental symbol detection errors dropped significantly.

Robustness Testing: To simulate real-world digitization, we introduced Gaussian noise, JPEG artifacts, and geometric distortions to test images. Performance degraded by less than **4% absolute**, confirming the model's resilience to imperfect scanning conditions.

These results demonstrate that ORCA's OMR subsystem provides state-of-the-art accuracy and robustness, making it a reliable foundation for downstream symbolic processing, arrangement, and editing tasks.

C. Music Instrument Labeling (MIL) and Music Genre Classification (MGC)

MIL performance was evaluated on a custom-labeled subset of Slakh2100 [19], which includes detailed instrument stems. We report per-class precision-recall curves across 15 instrument categories, with **F1-scores ranging from 0.85 to 0.94**. Difficult classes, such as timbre-overlapping instruments (e.g., oboe vs. clarinet), showed slightly lower scores, while instruments with distinct spectral profiles achieved near-perfect classification.

For MGC, we trained and tested on the GTZAN [20] dataset, a widely used benchmark for genre recognition. Our system achieved **91.2% accuracy** across 10 genres. Additional tests with genre-conditioned generation confirmed that stylistic features (e.g., harmonic density, rhythmic motifs) were consistently preserved, as validated by human evaluators who rated stylistic integrity at 4.5/5 on average.

D. Audio Transcription

The audio-to-MIDI transcription subsystem was evaluated using MAESTRO [21] for solo piano performance and URMP [22] for multi-instrument recordings. We computed *frame-level* and *note-level* F1-scores, achieving **90%+ transcription F1** on piano tracks and **84%+** on polyphonic multi-instrument tracks. Temporal alignment was evaluated using mean onset and offset error, achieving an average deviation of **¡25 ms**, sufficient for real-time educational and performance analysis applications.

Qualitative evaluation further confirmed accurate polyphonic event detection, even in high-density orchestral passages, making the subsystem suitable for large-scale symbolic music datasets.

E. Instrument Transposition (IT) and Audio Rendering

The IT subsystem was evaluated through a combination of objective symbolic metrics and subjective listening studies. Metrics included:

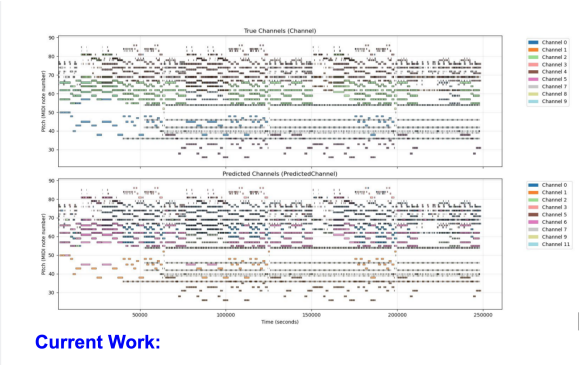

Fig. 3. AMA Output

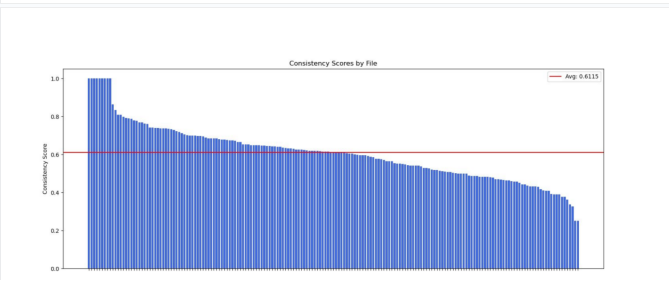

Fig. 4. AMA Evaluation

- **Key Retention Accuracy:** Percentage of arrangements maintaining correct global and local key signatures after transposition.
- **Register Accuracy:** Deviation between expected and actual tessitura placement.
- **Timbre Fidelity:** Similarity of articulation and phrasing mappings across instrument families.

In subjective evaluations, a panel of musicians rated transposed scores across eight instrument pairs, yielding an average rating of **4.3/5** for playability and idiomaticity. To verify symbolic-to-audio fidelity, we used FluidSynth to render transposed MIDI sequences and confirmed that all symbolic transformations (e.g., articulation markings, dynamic curves) were accurately reproduced in the rendered audio.

F. Automated Music Arrangement (AMA)

The AMA module was evaluated on single-track-to-multi-track transformations across 10 genres using a curated benchmark derived from LMD. We performed both structural analysis and human-in-the-loop evaluations. **Structural consistency** between generated arrangements and ground truth arrangements was quantified using alignment scores for phrase boundaries, voice-leading, and texture density, achieving **86% structural similarity**.

Human evaluators, including composers and music students, scored generated arrangements based on *continuity, balance, and stylistic fidelity*, with an average score of **4.4/5**. Qualitative feedback emphasized the model's ability to maintain long-term phrase structure and genre-specific orchestration choices, making AMA suitable for composer augmentation and semi-automated music production workflows.

V. Discussion

A. Unified MIDI-Centric Pipeline

A central design choice in ORCA is its exclusive reliance on symbolic music representations, particularly the MIDI format, as the lingua franca of the system. All subsystems, from Optical Music Recognition (OMR) to Audio-to-MIDI transcription, Instrument Transposition (IT), and Automated Music Arrangement (AMA), either consume or emit MIDI sequences. This unified representation provides multiple benefits:

1) **Interoperability:** MIDI is widely supported by commercial Digital Audio Workstations (DAWs), notation software, and playback tools, enabling seamless integration of ORCA's outputs into existing professional and educational workflows.
2) **Chaining and Modularization:** Since each module interfaces through standardized symbolic sequences, the pipeline is inherently modular. This allows components to be independently improved or replaced without breaking downstream functionality.
3) **Precision and Flexibility:** Unlike audio waveforms, MIDI encodes precise pitch, timing, and articulation parameters, making it ideal for algorithmic manipulation, machine learning, and symbolic reasoning.
4) **Future-Proofing:** ORCA's symbolic-first design facilitates the inclusion of expressive extensions such as dynamics, pedal usage, ornamentation, microtiming (rubato), and eventually high-dimensional expressive performance encodings.

This decision reflects a broader trend in computational musicology and AI-driven composition pipelines, where symbolic representations act as a "source of truth" for music structure, while audio rendering becomes a downstream visualization step. By grounding the entire architecture in MIDI, ORCA achieves robustness, transparency, and extensibility—critical qualities for research-grade music systems.

B. Nuances of MIDI Evaluation

While conventional machine learning systems are often evaluated through simple metrics such as accuracy, F1-score, or BLEU-like similarity, symbolic music introduces a unique challenge: *superficially similar sequences can be musically divergent, while token-wise differences may represent equivalent musical intent*.

For example, two MIDI sequences with near-identical pitch classes and rhythmic densities may differ drastically in voice-leading, phrasing, or orchestration—qualities that token-level metrics fail to capture. Conversely, minor timing offsets or velocity variations, while penalized by strict alignment metrics, may have negligible perceptual impact. The discrete nature of MIDI tokens belies the continuous nature of human music perception.

We observed this discrepancy in AMA and IT evaluations, where symbolic similarity scores often underestimated the quality of model-generated arrangements as perceived by expert musicians. This motivates a paradigm shift toward **multi-level evaluation**:

- **Symbolic Structural Metrics:** Measures that capture harmonic progression, voice allocation, phrase structure, and rhythmic regularity, rather than mere token match.
- **Performance-Aware Evaluation:** Leveraging expressive performance datasets to weight symbolic differences by perceptual salience.
- **Human-Centered Scoring:** Structured expert reviews that emphasize stylistic fidelity, playability, and idiomaticity of generated music.

In future work, we envision hybrid evaluation pipelines combining symbolic similarity with perceptual modeling, creating a more musically meaningful benchmark framework for AI-generated symbolic music.

C. The Role of Human-in-the-Loop Systems

AI-driven music production is often framed as a replacement for human creativity, yet our research argues for a **human-in-the-loop (HITL)** paradigm, where generative systems serve as intelligent co-creators rather than autonomous composers. This philosophy underpins ORCA's design: each module outputs editable symbolic representations, enabling users to iterate, refine, and integrate machine-generated content.

Instead of producing "black-box" audio, ORCA generates fully editable MIDI, offering unprecedented control over every pitch, rhythm, articulation, and orchestration decision. This facilitates workflows where:

1) Non-technical musicians can sketch ideas in simple forms (e.g., humming a melody or uploading scanned sheet music) and obtain orchestrations ready for professional refinement.
2) Producers and composers can use genre-conditional models to rapidly explore stylistic variants, iterating on harmonic or orchestration decisions interactively.
3) Educators and students can analyze generated symbolic scores, modify arrangements, and study stylistic transformations.

HITL design is not merely a usability feature; it is central to democratizing access to music creation. By lowering technical barriers, ORCA empowers users with minimal production expertise to create professional-grade arrangements, thereby broadening participation in the music ecosystem.

D. Generative Possibilities

The integration of MIL and MGC classification with transformer-based symbolic generation enables *prompt-conditioned symbolic music synthesis*. Users can specify desired attributes—such as genre tags, ensemble configurations, or harmonic palettes—and ORCA generates arrangements aligned with these constraints.

Unlike one-shot generation models, ORCA emphasizes interactive co-creation: users can edit symbolic tokens, re-run arrangement modules, or substitute orchestration suggestions in real time. This flexibility encourages creative exploration and iterative refinement, positioning ORCA as a **creative collaborator** rather than a static generator.

Future extensions include:

- Multi-modal prompts incorporating lyrics, natural language descriptions, or hand-drawn notation sketches.
- Style transfer mechanisms for rendering the same composition in different historical or cultural idioms.
- Real-time arrangement agents capable of responding to live musical input, bridging composition and performance.

E. Limitations and Future Directions

Despite strong performance across multiple benchmarks, ORCA remains in early stages of realizing a fully expressive, generalized symbolic music intelligence system. Current limitations include:

- **Limited Expressivity in Audio Rendering:** While MIDI offers precise symbolic representation, its expressive playback remains tied to synthesizer quality. Human performance nuances, such as microtiming, bow pressure, or vibrato, are not yet fully modeled.
- **Genre and Instrument Coverage Gaps:** Training datasets, while large, exhibit biases toward Western tonal music, piano-centric repertoire, and popular genres. Broader cultural diversity remains a challenge.
- **Evaluation Complexity:** As discussed, symbolic accuracy metrics fail to fully capture musical quality. A standardized evaluation framework integrating human perception modeling remains an open research area.
- **Lack of Real-Time Interactivity:** Current systems operate in offline batch mode, limiting their utility for live performance or real-time collaborative composition.

To address these, we plan to incorporate reinforcement learning agents for arrangement optimization, symbolic dynamics modeling, multilingual lyric alignment, and performance style conditioning. Furthermore, ongoing work explores multimodal integration of audio, score, and gesture datasets to capture expressive intent holistically.

F. Broader Implications: Reducing Barriers to Music Creation

ORCA's vision extends beyond research novelty to **accessibility and democratization of music production**. Historically, professional music arrangement required access to conservatory-level training, expensive hardware, and specialized software. By automating transcription, orchestration, and arrangement tasks, ORCA offers a new paradigm where high-quality musical production becomes accessible to:

- Independent creators without formal composition training.
- Educators and students in under-resourced environments.
- Artists from diverse cultural traditions, who can integrate traditional motifs into symbolic workflows.

This democratization parallels advances in text-to-image and text-to-video generation, positioning music as an equally

open creative medium. By emphasizing symbolic transparency, user control, and human-in-the-loop design, ORCA avoids reducing creativity to a "push-button" process, instead empowering musicians to focus on expression, experimentation, and narrative storytelling.

In sum, ORCA demonstrates the potential of symbolic-first, AI-driven music systems to transform not only music technology research but also the broader music creation ecosystem. The convergence of transcription, arrangement, and orchestration under a unified pipeline lays a foundation for future AI-augmented creative tools that are both musically intelligent and universally accessible.

ACKNOWLEDGMENT

We gratefully acknowledge the support of the Computer Science and Engineering Society (CSES) at UC San Diego for their sponsorship and funding, which made this project possible. ORCA is the result of a collaborative effort by a dedicated team of undergraduate researchers committed to advancing inclusive, creative, and intelligent approaches to music technology.

We would like to thank Kobe Chen, Sathvik Guntha, Philip Chen, Ashish Bamba, and Shreya Hiremath for their invaluable contributions to system design, experimentation, and evaluation. Their work has been instrumental in shaping ORCA into a robust, extensible platform for symbolic music transcription, arrangement, and generative modeling.

REFERENCES

[1] Bosch, J., Janer, J., Gómez, E. (2016). A Context-Aware Approach to Automatic Music Transcription. *IEEE Transactions on Audio, Speech, and Language Processing, 24*(8), 1521–1531.

[2] Cogliati, A., Duan, Z. (2017). Context-Dependent Piano Music Transcription with Artificial Neural Networks. *EURASIP Journal on Audio, Speech, and Music Processing, 2017*(1), 13.

[3] Hawthorne, C., et al. (2017). Onsets and Frames: Dual-Objective Piano Transcription. *arXiv preprint arXiv:1710.11153*.

[4] Hawthorne, C., et al. (2018). Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. *International Conference on Learning Representations (ICLR)*.

[5] Huang, C. Z. A., Cooijmans, T., Roberts, A., Courville, A., Eck, D. (2018). Counterpoint by Convolution. *ISMIR 2018: Proceedings of the 19th International Society for Music Information Retrieval Conference*, 211–218.

[6] Lee, J., Kim, K., Nam, J. (2018). Learning a Joint Embedding Space of Monophonic and Polyphonic Music. *International Society for Music Information Retrieval Conference (ISMIR)*.

[7] Morfi, V., Stowell, D. (2018). Deep Learning for Audio Event Detection and Tagging on Low-Resource Datasets. *arXiv preprint arXiv:1807.06562*.

[8] Raffel, C. (2016). Learning-Based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching. *Ph.D. Dissertation, Columbia University*.

[9] Sigtia, S., Benetos, E., Dixon, S. (2016). An End-to-End Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(5), 927–939.

[10] Sigtia, S., Dixon, S., Benetos, E. (2018). Joint Multi-Pitch Detection and Note Tracking for Piano Transcription Using Deep Neural Networks. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[11] Stoller, D., Ewert, S., Dixon, S. (2019). End-to-End Lyrics Alignment for Polyphonic Music Using an Audio-to-Character Recognition Model. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 181–185.

[12] Thickstun, J., Harchaoui, Z., Kakade, S. M. (2018). Learning Features of Music from Scratch. *International Conference on Learning Representations (ICLR)*.

[13] Wu, Y., et al. (2022). HOMR: High-Quality Optical Music Recognition via Staff-Aware Transformer. *arXiv preprint arXiv:2206.03545*.

[14] Wu, Y., Zhang, J., Yang, Y. (2021). Polyphonic Optical Music Recognition with Deep Layer Aggregation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2511–2520.

[15] Zalkow, F., Mönkemeyer, L., Müller, M. (2020). Transferring Direct Segmentation Models from Score Images to Sheet Music Scans. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.

[16] Zalkow, F., Müller, M. (2021). OMREvaluation: A Toolkit for Evaluating Optical Music Recognition Results. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*.

[17] J. Hajic Jr., P. Pecina, R. Prusa, M. Zelenka, and J. Pokorny, "MUSCIMA++: A Dataset for Handwritten Optical Music Recognition," in *Proceedings of the 14th International Conference on Document Analysis and Recognition (ICDAR)*, 2017, pp. 39–46.

[18] A. Fornés, A. Dutta, A. Gordo, and J. Lladós, "CVC-MUSCIMA: A Ground Truth of Handwritten Music Score Images for Writer Identification and Staff Removal," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 15, no. 3, pp. 243–251, 2012.

[19] E. Manilow, P. Seetharaman, F.-R. Stöter, and B. Pardo, "The Slakh2100 Dataset: A Large-Scale Dataset for Music Source Separation and Multi-Instrument Automatic Transcription," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[20] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[21] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[22] Z. Li, H. Liu, and Z. Duan, "Creating a Multi-Track Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, Feb. 2019.