SGCR: STRUCTURE-GRADIENT COUPLED REPLAY FOR CONTINUAL CRYSTAL PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Crystal structures and property distributions can evolve continuously as new material families and chemical spaces are explored, making continual learning essential for crystal property prediction. We study this problem under limited historical memory and propose SGCR, a structure-gradient coupled framework that jointly considers structural relevance and optimization interference. SGCR first constructs a frozen structural reference space to index a domain-balanced random memory of historical crystals. For each incoming batch, it retrieves historical samples according to both their structural relevance and their conflict with the current prediction objective, enabling more informative use of limited memory. We further introduce a calibrated structural-novelty signal that adaptively modulates gradient correction when current and historical objectives disagree, allowing the model to balance adaptation to newly arriving crystal distributions with retention of previously learned knowledge. Experiments across multiple datasets and backbones show that SGCR achieves lower average prediction error than a broad set of continual-learning baselines and alleviates catastrophic forgetting, with the size of the improvement varying across settings.

1 INTRODUCTION

Accurate crystal property prediction supports computational materials discovery by enabling rapid screening of candidate materials before costly experiments or first-principles calculations (Madika et al., 2025; Riebesell et al., 2025). Graph neural networks (GNNs) provide effective crystal representations by encoding atomic identities, local coordination environments, and geometric interactions in periodic structures (Xie & Grossman, 2018; Choudhary & DeCost, 2021; Yan et al., 2022; 2024). However, these models are typically developed under a static learning setting, in which training data are jointly accessible and drawn from a fixed distribution.

Materials discovery often proceeds sequentially, with new crystal data becoming available as different compositions and structural families are explored. The resulting distribution shifts require predictors to acquire new knowledge while maintaining accuracy on previously observed domains, as shown in Figure 1 (a). Continual learning addresses this requirement by learning from evolving data streams while mitigating catastrophic forgetting (Kirkpatrick et al., 2017; Chaudhry et al., 2019; Buzzega et al., 2020; Wang et al., 2024). Replay is a widely used strategy, including in recent online and graph continual learning methods, because a compact memory of historical examples provides direct supervision for retaining prior knowledge (Yoo et al., 2024; Su et al., 2025). Applying replay to continual crystal property prediction nevertheless raises two closely related challenges.

❶ **Replay must balance relevance to the current update with historical coverage.** A representative memory does not automatically identify which crystals matter for each update. Structural similarity captures relevance, while gradient conflict signals potential interference. Effective retrieval must combine these complementary cues while avoiding repeated concentration on a narrow subset of historical structures.

❷ **Conflict correction must balance knowledge retention with adaptation.** Excessive correction of conflicting gradients can hinder new learning, whereas insufficient correction can compromise historical predictions. Gradient disagreement captures local optimization tension but does not explicitly characterize structural distribution shifts. Structural novelty provides complementary con-

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

text for adapting correction strength, motivating a framework that couples structure-aware retrieval with adaptive gradient correction.

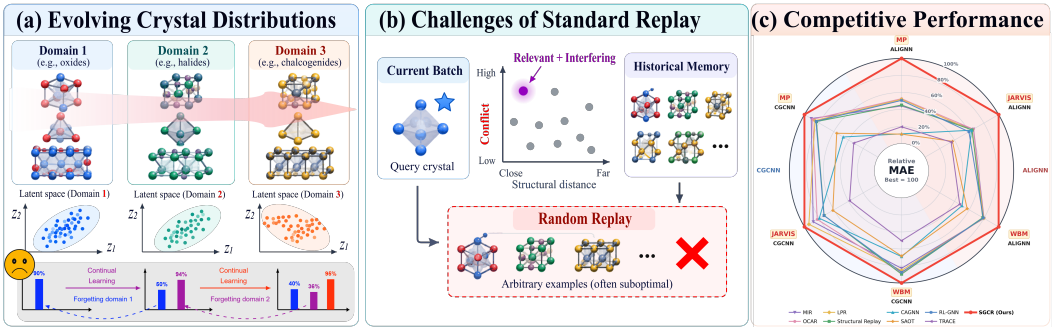

Figure 1: Challenges faced by traditional methods (a) and replay-based methods (b) under evolving crystal distributions, and the competitive performance achieved by our method.

In this paper, we propose SGCR, a structure-gradient coupled framework for continual crystal property prediction. SGCR first constructs a stable structural reference space to organize historical crystals and quantify the novelty of incoming structures. Using a domain-balanced random memory, it retrieves informative samples according to both structural relevance and their optimization conflict with the current data, complemented by sampling across other structural regions. Finally, the calibrated structural novelty adaptively controls gradient correction when current and historical objectives disagree, allowing the predictor to accommodate new crystal distributions while preserving previously learned knowledge. Our main contributions are summarized as follows:

- We formulate continual crystal property prediction under evolving crystal distributions and limited access to historical data, and propose SGCR to jointly support adaptation and knowledge retention.
- We develop a structure- and conflict-guided retrieval mechanism that selects replay samples from a domain-balanced random memory, combining local structural relevance and optimization conflict with sampling across other structural regions.
- We introduce calibrated structural novelty to adaptively modulate gradient correction, enabling update-specific control of interference between newly arriving and previously learned crystal distributions.
- Extensive experiments on three crystal datasets and three representative graph backbones show that SGCR outperforms strong continual-learning baselines on average, reducing MAE and RMSE by up to **47.33%** and **45.97%**, respectively, as shown in Figure 1(c).

2 Related Work

Crystal property prediction has increasingly shifted from handcrafted descriptors toward structure-aware neural representations. CGCNN (Xie & Grossman, 2018) models crystals as periodic atom graphs and directly learns structure–property relationships, while ALIGNN (Choudhary & DeCost, 2021) further incorporates bond-angle information through line-graph message passing. Subsequent methods place greater emphasis on periodicity and geometric expressiveness. Matformer (Yan et al., 2022) explicitly encodes periodic invariance, while ComFormer (Yan et al., 2024) and Crystalformer (Taniai et al., 2024) develop transformer architectures tailored to periodic crystal structures. More recent approaches further strengthen structural inductive biases through crystal-pattern representations (Gao et al., 2025), physicochemical attribute-aware graph encoding (Cao et al., 2026), and complementary invariant and equivariant geometric representations (Zhang et al., 2026a). Despite substantial progress on static crystal benchmarks, these methods primarily focus on property prediction under fixed training distributions, rather than sequential adaptation to evolving material domains. We discuss recent advances in continual learning in Appendix Section A.

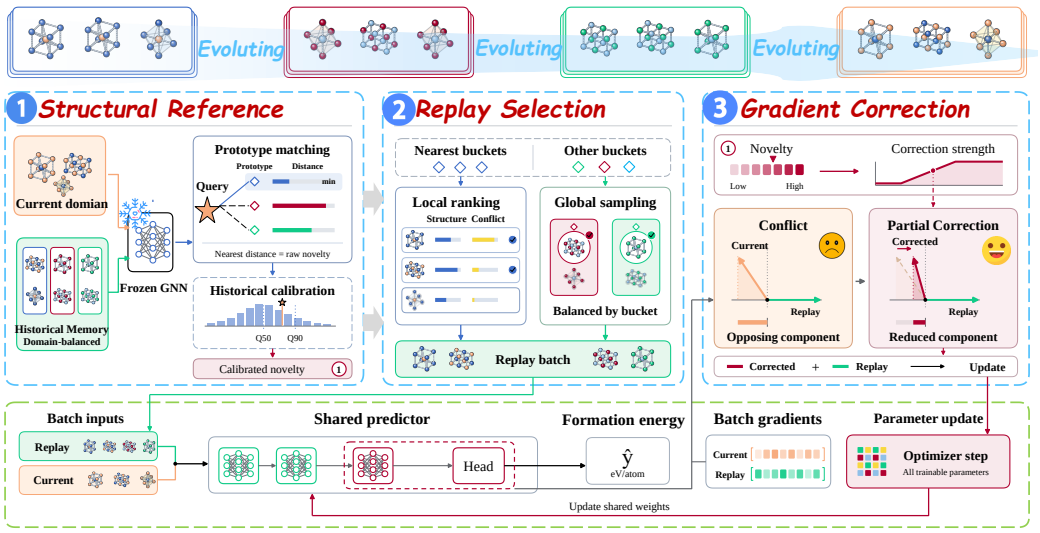

Figure 2: Overall framework of the model.

3 Problem Formulation

Crystal Inputs. A crystal is represented as $X = (\mathbf{Z}, \mathbf{S}, \mathbf{L})$, where $\mathbf{Z} \in \mathbb{N}^N$ denotes the atomic numbers of the N sites, $\mathbf{S} \in [0,1)^{N\times 3}$ their fractional coordinates, and $\mathbf{L} \in \mathbb{R}^{3\times 3}$ the lattice vectors. From X, we deterministically construct a periodic atom graph $G(X)$, where atoms are nodes and edges connect neighboring periodic images. Atomic descriptors and geometric features, such as interatomic distances, are derived from the crystal structure; angle-aware backbones additionally use a line graph to encode bond-angle information.

Crystal Property Prediction. Given a crystal structure X, or equivalently its deterministically constructed periodic graph $G(X)$, crystal property prediction aims to learn a parameterized mapping $f_\theta : X \to \hat{y}$, where $\hat{y} \in \mathbb{R}$ denotes the predicted crystal-level property and θ represents the learnable model parameters. Each crystal is associated with a scalar target $y \in \mathbb{R}$, such as formation energy per atom. Given a dataset $\mathcal{D} = \{(X_m, y_m)\}_{m=1}^{|\mathcal{D}|}$, the model is trained to predict $\hat{y}_m = f_\theta(X_m)$ so that the predicted properties closely match their corresponding ground-truth values.

Continual Crystal Property Prediction. Training proceeds through an ordered sequence of domains $\mathcal{D}_1, \ldots, \mathcal{D}_T$, whose crystal and property distributions may differ. At stage t, the learner receives the current training set $\mathcal{D}_t^{\mathrm{tr}}$ and a bounded memory $\mathcal{M}_{t-1} \subseteq \bigcup_{u<t} \mathcal{D}_u^{\mathrm{tr}}$, with $|\mathcal{M}_{t-1}| \le M$. A single predictor F_Θ is updated sequentially to learn the arriving domain while preserving predictive performance on previously observed domains. The goal is therefore to continually adapt to evolving crystal-property distributions under limited access to historical data.

4 Methodology

As shown in Figure 2 and Algorithm 1, SGCR combines stable structural relevance with gradient conflict to retrieve historical crystals, while sampling replay examples across other structural regions. It further uses structural novelty to modulate the correction of conflicting current gradients, coupling replay selection and model adaptation through a common structural reference.

4.1 Stable Structural Reference and Shared Prediction

Graph neural networks have emerged as effective encoders for crystalline materials, enabling crystal properties to be learned directly from atomic connectivity and local geometric environments (Xie & Grossman, 2018; Chen et al., 2019; Choudhary & DeCost, 2021). Building on this paradigm, we employ a shared crystal regressor that maps each periodic crystal graph to a fixed-dimensional

representation and subsequently predicts a scalar property:

$$\mathbf{h}_\theta(X) = f_\theta(G(X)) \in \mathbb{R}^d, \qquad F_\Theta(X) = \mathbf{w}^\top \mathbf{h}_\theta(X) + b. \tag{1}$$

where f_θ denotes the trainable GNN encoder applied to the periodic crystal graph $G(X)$, and $\mathbf{h}_\theta(X) \in \mathbb{R}^d$ is the resulting crystal-level representation. Depending on the backbone, f_θ performs message passing over atomic and geometric interactions followed by graph-level pooling. For example, we instantiate f_θ with ALIGNN (Choudhary & DeCost, 2021), which jointly propagates information over atom and line graphs to capture bond and angle interactions, or CGCNN (Xie & Grossman, 2018), which performs gated message passing over the crystal graph. The resulting representation is fed into a linear prediction head parameterized by $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, with $\Theta = (\theta, \mathbf{w}, b)$. SGCR operates on the crystal representation before this prediction head.

Structural measurements should remain comparable while the predictor adapts. After supervised learning on the first domain, we copy the selected encoder into $f_{\bar{\theta}}$ and freeze both its parameters and normalization statistics. The stable descriptor is $\mathbf{z}(X) = f_{\bar{\theta}}(G(X)) \in \mathbb{R}^d$. Although $f_{\bar{\theta}}$ originates from a learned encoder, its descriptors are fixed with respect to later optimization and may be cached. By contrast, $\mathbf{h}_\theta(X)$ changes as the predictor learns new domains. The stable descriptors define the geometry of historical memory; the predictive representations quantify the current model's disagreement with labeled examples.

4.2 Structural Indexing and Calibrated Novelty

At each domain boundary, we update the bounded memory $\mathcal{M}_t$ from $\mathcal{M}_{t-1} \cup \mathcal{D}_t^{\mathrm{tr}}$ by allocating the budget M evenly across observed domains and uniformly sampling without replacement within each domain. We organize the retained crystals into compact prototypes in the frozen reference space. These prototypes are used to measure and calibrate the structural novelty of incoming batches, providing a control signal for subsequent gradient correction.

Constructing the prototype index. A compact index organizes the stored descriptors into structural neighborhoods. Let $\bar{\mathbf{z}}$ and $\mathbf{v} \in \mathbb{R}^d_{\geq 0}$ be their coordinate-wise mean and population variance. We form at most K clusters from descriptors standardized by $(\mathbf{v} + \epsilon_s)^{1/2}$, using farthest-point initialization followed by Lloyd updates. Here and below, $\epsilon_s > 0$ stabilizes structural distances. Each nonempty cluster $\mathcal{P}_k$ defines a prototype with mean $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and diagonal population variance $\boldsymbol{\sigma}_k^2 \in \mathbb{R}^d_{\geq 0}$, both recomputed in the original descriptor coordinates. Prototype statistics are empirical quantities rather than learnable parameters. They are refreshed after memory updates and held fixed throughout the next stage.

Measuring novelty at the update scale. An update concerns a batch of arriving crystals, so its structural query is the mean stable descriptor $\mathbf{q} = |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \mathbf{z}(X_i)$ for the current batch $\mathcal{B}$. We measure its distance to prototype k and its historical novelty as

$$D_k(\mathbf{q}) = \left[\frac{1}{d} \sum_{a=1}^{d} \frac{(q_a - \mu_{k,a})^2}{\sigma_{k,a}^2 + \epsilon_s} \right]^{1/2}, \qquad \nu(\mathcal{B}) = \min_k D_k(\mathbf{q}). \tag{2}$$

To calibrate novelty, we form pseudo-batches from retained descriptors and compute their nearest-prototype distances. Let Q_{50} and Q_{90} denote the median and 90th percentile of the resulting historical novelty scores. The correction strength is then defined as

$$\rho(\mathcal{B}) = \alpha \, \mathrm{clip}_{[0,1]} \left(\frac{\nu(\mathcal{B}) - Q_{50}}{Q_{90} - Q_{50} + \epsilon_s} \right). \tag{3}$$

The coefficient $\alpha \in [0, 1]$ bounds the correction strength, while the clipped term calibrates current novelty against the historical distribution.

4.3 Structure and Conflict Guided Retrieval

Historical memory must support adaptation near the current structures while continuing to represent other regions. We rank nonempty replay-eligible prototype buckets by $D_k(\mathbf{q})$. Uniform sampling within buckets, balanced across prototypes, forms a local candidate set $\mathcal{C}_{\mathrm{loc}}$ from the nearest buckets

and a global candidate set $\mathcal{C}_{\text{glob}}$ from the remaining buckets. These are disjoint subsets of $\mathcal{M}^{\text{rep}}$. Separate candidate limits bound the cost of scoring, independently of the number of retained records.

Structural relevance provides a distance-based preference within the local candidates. For a historical candidate j, we use

$$d_j(\mathbf{q}) = \left[\frac{1}{d}\sum_{a=1}^{d}\frac{(z_{j,a}-q_a)^2}{v_a+\epsilon_s}\right]^{1/2}, \qquad r_j = \exp\left(-\frac{d_j(\mathbf{q})}{\tau}\right). \tag{4}$$

where $\mathbf{z}_j = \mathbf{z}(X_j)$ denotes the fixed structural descriptor of crystal X_j extracted by the frozen encoder. $\tau > 0$ is the median distance of sampled pairs of retained descriptors, lower-bounded by ϵ_s. Unlike the prototype distance in (Eq.2), this example-level distance uses the memory-wide variance $\mathbf{v}$, providing a common scale for comparing candidates from different buckets.

Structural relevance alone does not capture disagreement with the current objective. We estimate this conflict using a prediction-head signature. Let $\ell_\delta(e)$ denote the Huber loss with derivative $\psi_\delta(e) = \text{clip}(e, -\delta, \delta)$. For a labeled crystal, define

$$\mathbf{s}_i = \psi_\delta(F_\Theta(X_i) - y_i)\,\frac{\mathbf{h}_\theta(X_i)}{\max(\|\mathbf{h}_\theta(X_i)\|_2, \epsilon_n)}, \qquad \bar{\mathbf{s}}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbf{s}_i. \tag{5}$$

Both signatures lie in $\mathbb{R}^d$, and $\epsilon_n > 0$ stabilizes normalization. With residual $e_i = F_\Theta(X_i) - y_i$, the unnormalized product $\psi_\delta(e_i)\mathbf{h}_\theta(X_i)$ is the per-example gradient with respect to the linear head weights. The signature rescales the head-weight gradient by the feature norm and serves as a surrogate for gradient conflict. It is recomputed with the current predictor without gradient tracking, leaving model parameters and normalization statistics unchanged.

Then, we denote conflict score as $c_j = \max\{0, -\cos_{\epsilon_n}(\bar{\mathbf{s}}_{\mathcal{B}}, \mathbf{s}_j)\}$, where the stabilized cosine divides the inner product by $\max(\|\bar{\mathbf{s}}_{\mathcal{B}}\|_2, \epsilon_n)\max(\|\mathbf{s}_j\|_2, \epsilon_n)$ and is clipped to $[-1, 1]$. Large values indicate opposing signature directions. To combine relevance and conflict, each score is min–max normalized within the local candidate pool, with a numerically constant score mapped to zero. Local priority is then computed as

$$\pi_j = \lambda_s \widetilde{r}_j + \lambda_c \widetilde{c}_j, \qquad j \in \mathcal{C}_{\text{loc}}, \tag{6}$$

where $\lambda_s, \lambda_c \geq 0$ are fixed weights and the tildes denote normalized scores. A fixed fraction of replay samples is selected by descending π_j, while the remainder is drawn from $\mathcal{C}_{\text{glob}}$ with balanced prototype coverage, independent of conflict scores. Any allocation deficit is filled from the other pool without duplication, yielding a replay batch $\mathcal{R}$ with $|\mathcal{R}| = |\mathcal{B}|$.

4.4 Novelty-Modulated Gradient Correction

Retrieval determines which historical samples participate in an update, but the gradients induced by current and historical data may still conflict. We therefore jointly process the current batch $\mathcal{B}$ and the retrieved batch $\mathcal{R}$, while computing their mean Huber losses ℓ_δ separately:

$$\mathcal{L}_{\text{cur}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\ell_\delta(F_\Theta(X_i) - y_i), \qquad \mathcal{L}_{\text{rep}} = \frac{1}{|\mathcal{R}|}\sum_{j\in\mathcal{R}}\ell_\delta(F_\Theta(X_j) - y_j), \tag{7}$$
$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{cur}} + \lambda_r\mathcal{L}_{\text{rep}}.$$

The replay weight λ_r controls the contribution of historical supervision. Both losses supervise the shared predictor, while replay selection and structural novelty estimation are performed without gradient propagation.

Correction is restricted to a parameter subset $\Theta_{\mathcal{S}} \in \mathbb{R}^{p_{\mathcal{S}}}$ containing the final atom-graph message-passing block and scalar linear head. Earlier layers and any intermediate readout transformations remain trainable through the joint objective. We compute the actual restricted gradients $\mathbf{g}_{\text{cur}} = \nabla_{\Theta_{\mathcal{S}}}\mathcal{L}_{\text{cur}}$ and $\mathbf{g}_{\text{rep}} = \nabla_{\Theta_{\mathcal{S}}}\mathcal{L}_{\text{rep}}$, both in $\mathbb{R}^{p_{\mathcal{S}}}$. Unlike the retrieval signatures, these are obtained by differentiation through the current forward computation. Their inner product $\xi = \mathbf{g}_{\text{cur}}^\top\mathbf{g}_{\text{rep}}$ measures agreement on the selected parameters. The corrected current gradient is

$$\widetilde{\mathbf{g}}_{\text{cur}} = \begin{cases} \mathbf{g}_{\text{cur}} - \rho(\mathcal{B})\dfrac{\xi}{\|\mathbf{g}_{\text{rep}}\|_2^2 + \epsilon_g}\mathbf{g}_{\text{rep}}, & \text{if } \mathcal{A}, \\ \mathbf{g}_{\text{cur}}, & \text{otherwise,} \end{cases} \tag{8}$$

where $\epsilon_g > 0$ is a numerical safeguard and $\mathcal{A}$ requires $\xi < 0$, $\rho(\mathcal{B}) > 0$, and both squared gradient norms to exceed ϵ_g. The correction attenuates the component of the current gradient opposing the replay gradient. Its strength increases with calibrated structural novelty, up to α; it does not enforce an exact projection onto a nonconflicting half-space.

A backward pass through $\mathcal{L}_{\text{joint}}$ computes gradients for all trainable parameters. When $\mathcal{A}$ holds, the gradient on $\Theta_{\mathcal{S}}$ is replaced by $\widetilde{\mathbf{g}}_{\text{cur}} + \lambda_r \mathbf{g}_{\text{rep}}$, while all other gradients remain unchanged. We then apply global gradient-norm clipping followed by one AdamW step (Loshchilov & Hutter, 2019). Thus, structural novelty modulates the correction strength while preserving replay supervision.

4.5 Training Lifecycle and Inference

Training. First-domain learning optimizes F_Θ using supervised Huber regression and selects the checkpoint Θ_1 by validation MAE on $\mathcal{D}_1^{\text{val}}$. This checkpoint initializes continual training, while its encoder parameters are copied to $\bar{\theta}$ and frozen to define the structural reference $\mathbf{z}(X)$. A fresh optimizer is initialized after the first domain, with its state carried across subsequent stages. During stage $t \geq 2$, the historical memory $\mathcal{M}_{t-1}$ and prototype statistics $\{(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)\}_k$ remain fixed, while F_Θ supplies refreshed signatures $\mathbf{s}_i$ and gradients $\mathbf{g}_{\text{cur}}$ and $\mathbf{g}_{\text{rep}}$. At the stage boundary, domain-balanced random sampling updates $\mathcal{M}_t$ from $\mathcal{M}_{t-1} \cup \mathcal{D}_t^{\text{tr}}$, after which the structural index is rebuilt for the next domain. The final parameter state Θ_t is retained after each subsequent stage.

Inference. A crystal X is mapped to $\hat{y} = F_{\Theta_t}(X)$. Prediction uses only the learned regressor, without replay memory, the frozen encoder $f_{\bar{\theta}}$, prototype statistics, novelty estimates, or gradient signatures. The prediction architecture therefore remains fixed as domains accumulate.

5 Experiment

5.1 Experiment Setup

Datasets and Continual Learning Setup We evaluate formation-energy prediction (eV/atom) on three datasets. ❶ **MP** (Jain et al., 2013) contains 132,752 crystals, partitioned into seven crystal-system domains and presented sequentially as triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic. ❷ **JARVIS-DFT** (Choudhary et al., 2020) comprises 93,902 materials from the 2021, 2022, and 2025 snapshots. Three chronological domains are constructed by assigning each material to its first valid appearance while retaining its corresponding structure and label. ❸ **WBM** (Wang et al., 2021) contains 236,624 records after structural deduplication, retaining the earliest occurrence of each duplicate group. Its five chemical-substitution rounds define five ordered domains, using official relaxed structures and formation energies, as shown in Appendix B.1.

Setting Experiments are conducted on a single NVIDIA GeForce RTX 5090 GPU with 32 GB of memory. All methods are implemented in PyTorch using ALIGNN (Choudhary & DeCost, 2021), CGCNN (Xie & Grossman, 2018), or Matformer (Yan et al., 2022) as the backbone, with identical data partitions and domain orders for fair comparison. Models are optimized with AdamW using a weight decay of 10^{-4} and Huber loss. The learning rate is set to 10^{-3} for the first domain and 5×10^{-4} for subsequent domains, with a current-batch size of 32 and global gradient-norm clipping at 5. *Replay-based methods use a total memory budget of 500 historical training samples*. For SGCR, the memory is populated by domain-balanced random sampling; 400 samples are used for replay and 100 are reserved for structural statistics, with each current batch paired with a replay batch of equal size. The first domain is trained for up to 20 epochs with validation-based early stopping of patience 5. Each subsequent domain is trained for 10 epochs. We run our model three times and report the mean and standard deviation across the three runs.

Evaluation Metrics. We report final MAE and RMSE by equally averaging domain-wise test errors after learning the last domain, and use "Forgetting" to measure knowledge retention. Detailed definitions are provided in Appendix B.3.

Baselines We compare SGCR with a broad set of continual-learning baselines, including Direct which denotes a sequentially fine-tuned model without any continual learning strategy; SI (Zenke

et al., 2017), MAS (Aljundi et al., 2018), ER (Rolnick et al., 2019), A-GEM (Chaudhry et al., 2019), MIR (Aljundi et al., 2019), Online EWC (Schwarz et al., 2018), CAGNN (Xiao et al., 2026), SAOT (Zhang et al., 2026b), RL-GNN (Song et al., 2025), HPNs (Zhang et al., 2023), OCAR (Uretini & Carta, 2025), TRACE (Peng et al., 2025), and LPR (Yoo et al., 2024). For a fair comparison, all methods follow the same data-partitioning protocol and use the same crystal backbone initialized from the first-domain checkpoint and follow an identical continual-learning sequence.

Table 1: Performance comparison on MP, JARVIS, and WBM with ALIGNN and CGCNN as backbones. MAE and RMSE are final domain-macro averages in eV/atom. SGCR results are reported as mean ± standard deviation over three runs. For each baseline, the orange percentage indicates its relative error increase over SGCR under the same backbone. The **best** results and second-best results are highlighted.

Type	Method	MP		JARVIS		WBM		Average	
		MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
ALIGNN backbone [npj Comput. Mater. 2021]									
General CL	Direct	0.2229 ↑272.74%	0.3519 ↑277.58%	0.1092 ↑113.70%	0.1677 ↑102.05%	0.1192 ↑24.69%	0.7388 ↑14.72%	0.1504 ↑118.60%	0.4195 ↑53.44%
	SI [ICML 2017]	0.2495 ↑317.22%	0.3779 ↑305.47%	0.1012 ↑98.04%	0.1439 ↑73.37%	0.1610 ↑68.41%	0.7516 ↑16.71%	0.1706 ↑147.97%	0.4245 ↑55.27%
	MAS [ECCV 2018]	0.4883 ↑716.59%	0.5686 ↑510.09%	0.1111 ↑117.42%	0.1513 ↑82.29%	0.1847 ↑93.20%	0.7619 ↑18.31%	0.2614 ↑279.94%	0.4939 ↑80.65%
	ER [NeurIPS 2019]	0.3212 ↑437.12%	0.4295 ↑360.84%	0.0815 ↑59.49%	0.1292 ↑55.66%	0.1222 ↑27.82%	0.7396 ↑14.84%	0.1750 ↑154.36%	0.4328 ↑58.30%
	A-GEM [ICLR 2019]	0.2028 ↑239.13%	0.3029 ↑225.00%	0.1004 ↑96.48%	0.1552 ↑86.99%	0.1215 ↑27.09%	0.7387 ↑14.70%	0.1416 ↑105.81%	0.3989 ↑45.90%
	MIR [NeurIPS 2019]	0.1317 ↑120.23%	0.2370 ↑154.29%	0.0763 ↑49.32%	0.1218 ↑46.75%	0.1140 ↑19.25%	0.7326 ↑13.76%	0.1073 ↑55.96%	0.3638 ↑33.07%
	Online EWC [ICML 2018]	0.0746 ↑24.75%	0.0982 ↑5.36%	0.0624 ↑22.11%	0.0912 ↑9.88%	0.1024 ↑7.11%	0.6914 ↑7.36%	0.0798 ↑15.99%	0.2936 ↑7.39%
	OCAR [ICML 2025]	0.1103 ↑84.45%	0.1737 ↑86.37%	0.0766 ↑49.90%	0.1144 ↑37.83%	0.1150 ↑20.29%	0.7350 ↑14.13%	0.1006 ↑46.22%	0.3410 ↑24.73%
	LPR [ICML 2024]	0.0610 ↑2.01%	0.0989 ↑6.12%	0.0566 ↑10.76%	0.0855 ↑3.01%	0.0991 ↑3.66%	0.6804 ↑5.65%	0.0722 ↑4.94%	0.2883 ↑5.45%
Graph CL	HPNs [TPAMI 2023]	0.1284 ↑114.72%	0.2058 ↑120.82%	0.0787 ↑54.01%	0.1188 ↑43.13%	0.1165 ↑21.86%	0.6612 ↑2.67%	0.1079 ↑56.83%	0.3286 ↑20.19%
	CAGNN [ICML 2026]	0.5128 ↑757.53%	0.5970 ↑540.56%	0.0795 ↑55.58%	0.1200 ↑64.58%	0.1815 ↑89.85%	0.7038 ↑9.29%	0.2579 ↑274.85%	0.4736 ↑73.23%
	SAOT [ICML 2026]	0.5274 ↑781.94%	0.6140 ↑558.80%	0.1353 ↑164.77%	0.1794 ↑116.14%	0.1596 ↑66.95%	0.7585 ↑17.78%	0.2741 ↑298.40%	0.5173 ↑89.21%
	RL-GNN [AAAI 2025]	0.0601 ↑0.50%	0.0981 ↑5.26%	0.0541 ↑5.87%	0.0899 ↑8.31%	0.0983 ↑2.82%	0.6969 ↑8.21%	0.0709 ↑3.05%	0.2950 ↑7.90%
	TRACE [KDD 2025]	0.2963 ↑395.48%	0.4011 ↑330.36%	0.1421 ↑178.08%	0.1883 ↑126.87%	0.1939 ↑102.82%	0.7999 ↑24.21%	0.2108 ↑206.40%	0.4631 ↑69.39%
SGCR (Ours)		**0.0598 ± 0.0053**	**0.0932 ± 0.0039**	**0.0511 ± 0.0007**	**0.0830 ± 0.0010**	**0.0956 ± 0.0060**	**0.6440 ± 0.0635**	**0.0688**	**0.2734**
CGCNN backbone [PRL 2018]									
General CL	Direct	0.1206 ↑60.37%	0.1674 ↑46.97%	0.1162 ↑57.88%	0.1922 ↑70.69%	0.1226 ↑29.60%	0.7451 ↑18.03%	0.1198 ↑47.72%	0.3682 ↑28.79%
	SI [ICML 2017]	0.2032 ↑170.21%	0.2862 ↑151.27%	0.1070 ↑45.38%	0.1484 ↑31.79%	0.1625 ↑71.78%	0.7397 ↑17.17%	0.1576 ↑94.33%	0.3914 ↑36.90%
	MAS [ECCV 2018]	0.1618 ↑115.16%	0.2498 ↑119.32%	0.1176 ↑59.78%	0.1543 ↑37.03%	0.1694 ↑79.07%	0.7426 ↑17.63%	0.1496 ↑84.46%	0.3822 ↑33.68%
	ER [NeurIPS 2019]	0.0855 ↑13.70%	0.1253 ↑10.01%	0.0876 ↑19.02%	0.1359 ↑20.69%	0.1185 ↑25.26%	0.7357 ↑16.54%	0.0972 ↑19.85%	0.3323 ↑16.23%
	A-GEM [ICLR 2019]	0.6642 ↑783.24%	0.7178 ↑530.20%	0.1800 ↑144.57%	0.2577 ↑128.86%	0.2047 ↑116.38%	0.7789 ↑23.38%	0.3496 ↑331.07%	0.5848 ↑104.55%
	MIR [NeurIPS 2019]	0.0810 ↑7.71%	0.1240 ↑8.87%	0.0829 ↑12.64%	0.1261 ↑11.09%	0.1306 ↑38.05%	0.7507 ↑18.91%	0.0982 ↑21.09%	0.3336 ↑16.68%
	Online EWC [ICML 2018]	0.1045 ↑58.96%	0.1786 ↑56.80%	0.1110 ↑50.82%	0.1627 ↑44.49%	0.1211 ↑28.01%	0.7446 ↑17.95%	0.1122 ↑38.35%	0.3620 ↑26.62%
	OCAR [ICML 2025]	0.0827 ↑9.97%	0.1255 ↑10.18%	0.0884 ↑20.11%	0.1202 ↑6.75%	0.1226 ↑29.60%	0.7441 ↑17.87%	0.0979 ↑20.72%	0.3299 ↑15.39%
	LPR [ICML 2024]	0.0858 ↑14.10%	0.1290 ↑13.26%	0.0777 ↑5.57%	0.1244 ↑10.48%	0.1267 ↑33.93%	0.7381 ↑16.92%	0.0967 ↑19.24%	0.3305 ↑15.60%
Graph CL	HPNs [TPAMI 2023]	0.0879 ↑16.89%	0.1331 ↑16.86%	0.0997 ↑35.46%	0.1344 ↑19.36%	0.1206 ↑27.48%	0.7461 ↑18.18%	0.1027 ↑26.63%	0.3379 ↑18.19%
	CAGNN [ICML 2026]	0.1553 ↑106.52%	0.2603 ↑128.53%	0.0915 ↑24.32%	0.1356 ↑20.43%	0.1566 ↑65.54%	0.7508 ↑18.93%	0.1345 ↑65.84%	0.3822 ↑33.68%
	SAOT [ICML 2026]	0.1311 ↑74.34%	0.2009 ↑76.38%	0.1189 ↑61.55%	0.1728 ↑53.46%	0.1583 ↑67.34%	0.7574 ↑19.97%	0.1361 ↑67.82%	0.3770 ↑31.86%
	RL-GNN [AAAI 2025]	0.0872 ↑15.96%	0.1367 ↑20.02%	0.0994 ↑35.05%	0.1492 ↑32.50%	0.1237 ↑30.76%	0.7528 ↑19.25%	0.1034 ↑27.50%	0.3462 ↑21.09%
	TRACE [KDD 2025]	0.2227 ↑196.14%	0.3048 ↑167.60%	0.1890 ↑156.79%	0.2705 ↑140.23%	0.2150 ↑127.27%	0.8179 ↑29.56%	0.2089 ↑157.58%	0.4644 ↑62.43%
SGCR (Ours)		**0.0752 ± 0.0031**	**0.1139 ± 0.0039**	**0.0736 ± 0.0007**	**0.1126 ± 0.0029**	**0.0946 ± 0.0125**	**0.6313 ± 0.0917**	**0.0811**	**0.2859**

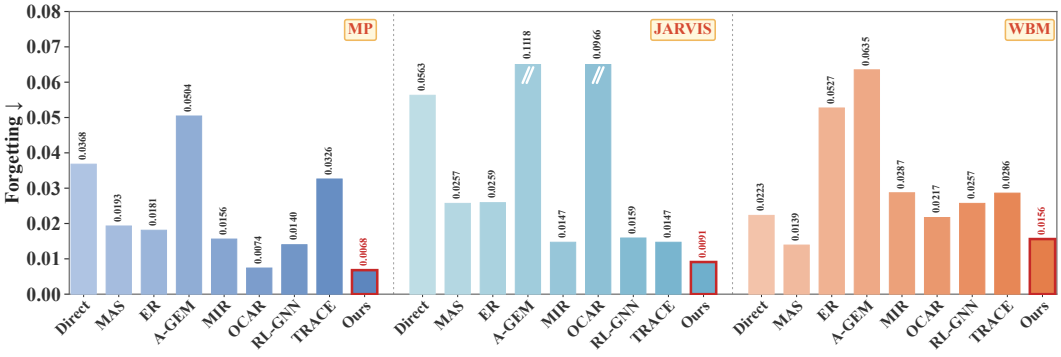

Figure 3: Forgetting comparison using ALIGNN as the backbone.

5.2 Main Results

Table 1 and Figure 3 show that replay-based methods are generally competitive across the three datasets, highlighting the importance of historical supervision. MIR and OCAR perform well in several settings, and HPNs is competitive on WBM and average RMSE. Graph-specific methods behave less consistently: RL-GNN remains competitive in several settings, whereas CAGNN, SAOT, and TRACE degrade notably on some dataset–backbone combinations, suggesting that graph-specific strategies do not necessarily transfer to continual crystal property regression, where both structural distributions and property mappings evolve.

SGCR attains the lowest mean error in all **12** settings (two backbones, three datasets, two metrics), but the margins vary. They are largest with CGCNN, e.g., 20.17% lower MAE than the strongest baseline on WBM, whereas for ALIGNN on MP MAE and WBM RMSE the gap is within the standard deviation of SGCR, so the methods should be regarded as comparable there.

5.3 Ablation Study

We conduct ablation studies on MP and WBM with ALIGNN and CGCNN backbones to examine the contribution of each component in SGCR. **w/o Conflict Scoring** removes the conflict term from replay prioritization and selects local historical samples solely according to structural relevance. **w/o Structural Retrieval** removes structure-guided replay selection, thereby discarding the structural correspondence between incoming and historical crystals. **w/o Gradient Correction** disables the correction of conflicting current gradients while retaining replay-based joint optimization. **Fixed Correction** replaces the novelty-dependent correction strength $\rho(\mathcal{B})$ with a constant coefficient, removing the adaptive modulation provided by structural novelty.

As shown in Figure 4, removing these components generally leads to higher final MAE across datasets and backbones, although the differences are modest in several settings. Without conflict scoring, replay selection cannot explicitly prioritize historical samples that are likely to interfere with the current update, which may reduce the utility of limited replay capacity. Removing structural retrieval weakens the correspondence between incoming crystals and relevant historical regions, so replay samples may provide less targeted supervision under evolving structural distributions. Disabling gradient correction also increases error, although by a smaller margin, indicating that correction complements rather than replaces structure- and conflict-aware replay. Finally, a fixed correction strength does not account for variation in structural shift across updates and applies the same preservation intensity regardless of how novel the incoming structures are.

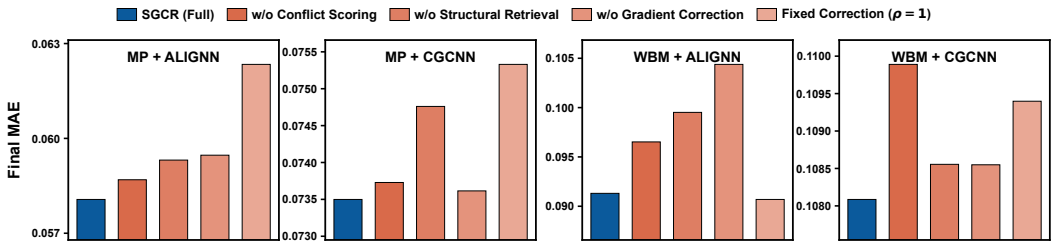

Figure 4: Ablation Study of SGCR.

5.4 Effect of Novelty-Dependent Gradient Correction

We examine whether structural novelty provides useful guidance for gradient correction. We compare SGCR with no correction, fixed correction, gradient-only control, and reverse- and permuted-novelty controls. The coefficient ρ denotes the correction strength: no correction sets $\rho = 0$, while the fixed variant uses $\rho = 1$. Gradient-only control sets $\rho = 0.1 \max\{0, -\cos(\mathbf{g}_{\text{cur}}, \mathbf{g}_{\text{rep}})\}$, where $\mathbf{g}_{\text{cur}}$ and $\mathbf{g}_{\text{rep}}$ are the current and replay gradients on the corrected parameter subset. Reverse and permuted controls preserve SGCR's coefficient distribution but respectively reverse or shuffle its correspondence with structural novelty.

As shown in Figure 5, removing correction causes only a small degradation, whereas all misaligned variants perform worse than no correction, indicating that gradient correction is double-edged: correction that is not aligned with structural novelty can harm retention. The fixed variant ($\rho = 1$) corresponds to a full A-GEM-style projection and serves as an upper-strength reference. More importantly, the permuted-novelty control preserves the exact distribution, and hence the average magnitude, of SGCR's coefficients while breaking their correspondence with incoming structures; it yields the largest degradation, suggesting that the benefit arises from aligning correction strength with structural novelty rather than from its magnitude. We therefore view novelty-modulated correction as an incremental refinement within SGCR.

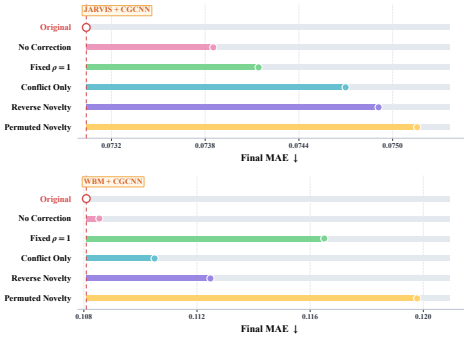

Figure 5: Sensitivity analysis of gradient-correction control.

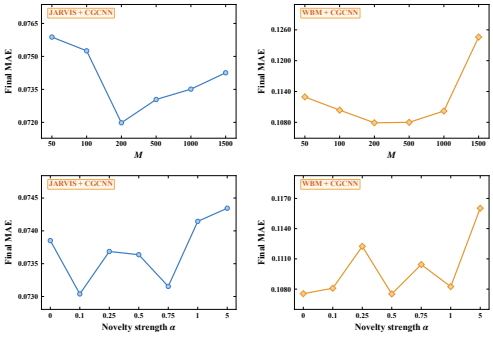

Figure 6: Hyperparameter sensitivity analysis.

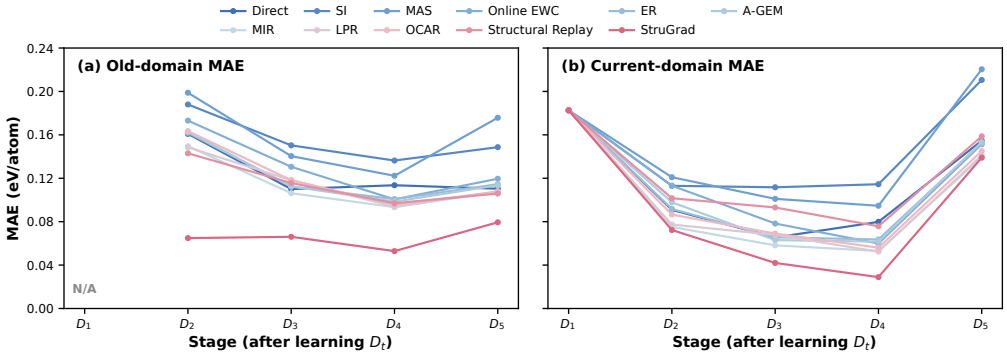

Figure 7: Stability–plasticity analysis on WBM with ALIGNN.

5.5 Hyperparameter Sensitivity

We examine the sensitivity of SGCR to the memory budget M and novelty strength α using CGCNN on JARVIS and WBM. As shown in Figure 6, on both datasets, increasing the memory budget initially reduces final MAE, with intermediate budgets providing the most favorable accuracy. Further expansion offers limited benefit and eventually increases prediction error, particularly on WBM, indicating that a larger historical memory does not necessarily improve continual prediction. The effect of novelty strength is also non-monotonic and dataset-dependent: JARVIS favors relatively weak correction, whereas WBM achieves its lowest error at an intermediate strength. Strong correction degrades accuracy on both datasets, highlighting the importance of controlling the magnitude of gradient modification.

5.6 Stability–Plasticity Analysis

We examine whether the improved retention of SGCR comes at the cost of adaptation. Figure 7 reports the average MAE over previously observed domains and the current-domain MAE after each stage. SGCR achieves lower historical-domain error than the compared methods while also yielding lower current-domain MAE, suggesting that, on WBM with ALIGNN, its stability is not obtained by suppressing adaptation.

6 Conclusion

We study continual crystal property prediction under evolving structural distributions and limited historical memory, and propose SGCR, which couples structure- and conflict-aware replay with novelty-modulated gradient correction. Experiments across multiple datasets and backbones show improved average accuracy and knowledge retention over a range of continual-learning baselines, with gains that vary across settings.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and polishing during manuscript preparation, as well as to provide suggestions for improving the visual presentation of the model framework in Figure 2. The final figure was manually created and refined by the authors using Microsoft PowerPoint.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we make the source code for our model publicly available in an anonymous repository.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision*, pp. 139–154, 2018. doi: 10.1007/978-3-030-01219-9_9.

Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11849–11860, 2019.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jie Cao, Kai Huang, Jian Mao, Mengya Zhong, Chen Chen, Kexun Li, and Taikang Liu. Physics-informed graph neural network representation learning for crystal property prediction. *npj Computational Materials*, 2026. doi: 10.1038/s41524-026-02131-9.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*, 2019.

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019. doi: 10.1021/acs.chemmater.9b01294.

Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7:185, 2021. doi: 10.1038/s41524-021-00650-1. URL https://doi.org/10.1038/s41524-021-00650-1.

Kamal Choudhary, Kevin F. Garrity, Andrew C. E. Reid, Brian DeCost, Adam J. Biacchi, Angela R. Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A. Gilad Kusne, Andrea Centrone, Albert Davydov, Jie Jiang, Ruth Pachter, Gowoon Cheon, Evan Reed, Ankit Agrawal, Xiaofeng Qian, Vinit Sharma, Houlong Zhuang, Sergei V. Kalinin, Bobby G. Sumpter, Ghanshyam Pilania, Pinar Acar, Subhasish Mandal, Kristjan Haule, David Vanderbilt, Karin Rabe, and Francesca Tavazza. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6(1):173, 2020. doi: 10.1038/s41524-020-00440-1.

Hengda Gao, Xiao-Wei Guo, Genglin Li, Chao Li, and Canqun Yang. Gcpnet: An interpretable generic crystal pattern graph neural network for predicting material properties. *Neural Networks*, 188:107466, 2025. doi: 10.1016/j.neunet.2025.107466.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting

in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Huihui Liu, Yiding Yang, and Xinchao Wang. Overcoming catastrophic forgetting in graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8653–8661, 2021. doi: 10.1609/aaai.v35i10.17049.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6467–6476, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Benediktus Madika, Aditi Saha, Chaeyul Kang, Batzorig Buyantogtokh, Joshua Agar, Chris M. Wolverton, Peter Voorhees, Peter Littlewood, Sergei Kalinin, and Seungbum Hong. Artificial intelligence for materials discovery, development, and optimization. *ACS Nano*, 19(30):27116–27158, 2025. doi: 10.1021/acsnano.5c04200.

Zhen Peng, Xu Hua, Jingchen Hao, Qika Lin, Bo Dong, and Chao Shen. Self-supervised continual graph learning via adaptive spaced replay on node proxies. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 2303–2314, 2025. doi: 10.1145/3711896.3737121.

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Chiang Yuan, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7:836–847, 2025. doi: 10.1038/s42256-025-01055-1.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 348–358, 2019.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537. PMLR, 2018.

Lei Song, Jiaxing Li, Qinghua Si, Shihan Guan, and Youyong Kong. Exploring rationale learning for continual graph learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20540–20548, 2025. doi: 10.1609/aaai.v39i19.34263.

Junwei Su, Difan Zou, and Chuan Wu. On the limitation and experience replay for gnns in continual learning. In *Proceedings of the 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 342–366. PMLR, 2025.

Tatsunori Taniai, Ryo Igarashi, Yuta Suzuki, Naoya Chiba, Kotaro Saito, Yoshitaka Ushiku, and Kanta Ono. Crystalformer: Infinitely connected attention for periodic structure encoding. In *International Conference on Learning Representations*, 2024.

Ruilin Tong, Yuhang Liu, Javen Qinfeng Shi, and Dong Gong. Coreset selection via reducible loss in continual learning. In *International Conference on Learning Representations*, 2025.

Edoardo Urettini and Antonio Carta. Online curvature-aware replay: Leveraging $\mathbf{2^{nd}}$ order information for online continual learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60590–60609. PMLR, 2025.

Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, 7(1):12, 2021. doi: 10.1038/s41524-020-00481-6.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329.

Ye Xiao, Ruikun Li, Zhenyu Yang, Andrey Vasnev, and Junbin Gao. Capacity-agnostic parameter isolation for continual graph learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=hkcF6wKqoq.

Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14): 145301, 2018. doi: 10.1103/PhysRevLett.120.145301. URL https://doi.org/10.1103/PhysRevLett.120.145301.

Keqiang Yan, Yi Liu, Yuchao Lin, and Shuiwang Ji. Periodic graph transformers for crystal material property prediction. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15066–15080, 2022.

Keqiang Yan, Cong Fu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Complete and efficient graph transformers for crystal material property prediction. In *International Conference on Learning Representations*, 2024.

Jinsoo Yoo, Yunpeng Liu, Frank Wood, and Geoff Pleiss. Layerwise proximal replay: A proximal point method for online continual learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57199–57216. PMLR, 2024.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017.

Liang Zhang, Ziyue Wang, Xin Wang, Xinyu Zhang, Yu Mao, Kong Chen, and Yuen Wu. Improving crystal material property prediction with multi-view geometric graph transformer. *Nature Communications*, 17(1):7148, 2026a. doi: 10.1038/s41467-026-73627-7.

Xikun Zhang, Dongjin Song, and Dacheng Tao. Hierarchical prototype networks for continual graph representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4622–4636, 2023. doi: 10.1109/TPAMI.2022.3186909.

Yuting Zhang, Yanbei Liu, Zhitao Xiao, Lei Geng, Yanwei Pang, and Xiao Wang. SAOT: Self-supervised continual graph learning with structure-aware optimal transport, 2026b.

Fan Zhou and Chengtai Cao. Overcoming catastrophic forgetting in graph neural networks with experience replay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4714–4722, 2021. doi: 10.1609/aaai.v35i5.16602.

A Related Work

Continual Learning. Continual learning aims to acquire knowledge from sequentially arriving data while mitigating catastrophic forgetting (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019; Buzzega et al., 2020; Wang et al., 2024). Two major families relevant to our setting are replay-based and parameter-regularization methods. Replay-based approaches retain a compact subset of historical samples and revisit them during subsequent learning, with recent studies improving memory selection through interference-aware retrieval (Aljundi et al., 2019) or informative coreset construction (Tong et al., 2025). Parameter-regularization methods instead restrict changes to parameters important for previously learned knowledge, as exemplified by EWC (Kirkpatrick et al., 2017) and Synaptic Intelligence (Zenke et al., 2017). Continual learning has also been extended to graph-structured data. ER-GNN (Zhou & Cao, 2021) preserves representative historical nodes through experience replay, whereas TWP (Liu et al., 2021) protects parameters important to both task prediction and graph topology. These methods mainly target generic graph learning settings such as node or graph classification. In contrast, SGCR addresses continual crystal property regression by jointly exploiting crystal structural representations for historical memory organization and adaptive knowledge preservation.

Algorithm 1 Continual Training of SGCR

Input: Ordered domains $\{\mathcal{D}_t\}_{t=1}^{T}$, memory budget M, prototype number K
Output: Continually updated predictor F_{Θ_T}

1. First-Domain Initialization

Train F_Θ on $\mathcal{D}_1^{\mathrm{tr}}$ with Huber regression
Select Θ_1 using validation MAE
Freeze the first-domain encoder $f_{\bar{\theta}}$ and define $\mathbf{z}(X) \leftarrow f_{\bar{\theta}}(G(X))$
Initialize historical memory $\mathcal{M}_1$
Build the structural prototype index and calibrate historical novelty

2. Structural Novelty Estimation

for $t = 2$ **to** T **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}_t^{\mathrm{tr}}$ **do**
    Compute the batch structural query: $\mathbf{q} \leftarrow \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbf{z}(X_i)$
    Compute historical novelty: $\nu(\mathcal{B}) \leftarrow \min_k D_k(\mathbf{q})$
    Calibrate the correction strength: $\rho(\mathcal{B}) \leftarrow \alpha\,\mathrm{clip}_{[0,1]}\left(\frac{\nu(\mathcal{B})-Q_{50}}{Q_{90}-Q_{50}+\epsilon_s}\right)$

3. Structure–Conflict Guided Retrieval

    Retrieve nearby prototype buckets as local candidates $\mathcal{C}_{\mathrm{loc}}$
    Sample remaining prototype regions as global candidates $\mathcal{C}_{\mathrm{glob}}$
    Evaluate structural relevance r_j and prediction conflict c_j for local candidates
    Compute local retrieval priority: $\pi_j \leftarrow \lambda_s \widetilde{r}_j + \lambda_c \widetilde{c}_j$
    Select high-priority local samples and balanced global samples
    Form replay batch $\mathcal{R}$ with $|\mathcal{R}| = |\mathcal{B}|$

4. Novelty-Modulated Gradient Update

    Compute current and replay losses $\mathcal{L}_{\mathrm{cur}}$ and $\mathcal{L}_{\mathrm{rep}}$
    Form the joint objective: $\mathcal{L}_{\mathrm{joint}} \leftarrow \mathcal{L}_{\mathrm{cur}} + \lambda_r \mathcal{L}_{\mathrm{rep}}$
    Compute restricted gradients $\mathbf{g}_{\mathrm{cur}}$ and $\mathbf{g}_{\mathrm{rep}}$ on $\Theta_{\mathcal{S}}$
    $\xi \leftarrow \mathbf{g}_{\mathrm{cur}}^{\top}\mathbf{g}_{\mathrm{rep}}$
    if $\xi < 0$, $\rho(\mathcal{B}) > 0$, **and** $\min(\|\mathbf{g}_{\mathrm{cur}}\|_2^2, \|\mathbf{g}_{\mathrm{rep}}\|_2^2) > \epsilon_g$ **then**
      $\widetilde{\mathbf{g}}_{\mathrm{cur}} \leftarrow \mathbf{g}_{\mathrm{cur}} - \rho(\mathcal{B})\frac{\xi}{\|\mathbf{g}_{\mathrm{rep}}\|_2^2+\epsilon_g}\mathbf{g}_{\mathrm{rep}}$
    else
      $\widetilde{\mathbf{g}}_{\mathrm{cur}} \leftarrow \mathbf{g}_{\mathrm{cur}}$
    end if
    Replace the restricted gradient by $\widetilde{\mathbf{g}}_{\mathrm{cur}} + \lambda_r \mathbf{g}_{\mathrm{rep}}$
    Apply global gradient clipping and update F_Θ
  end for

5. Memory and Index Refresh

  Update $\mathcal{M}_t$ by domain-balanced random sampling
  Rebuild the prototype index from $\mathcal{M}_t$
  Recalibrate Q_{50} and Q_{90} for the next domain
end for
return F_{Θ_T}

B Experimental Details

B.1 Dataset Details

We provide additional details on the three datasets used for continual formation-energy prediction. The datasets represent three complementary forms of distribution evolution in crystalline materials: crystallographic variation across crystal systems in Materials Project (MP), chronological database expansion in JARVIS-DFT, and successive chemical-substitution rounds in WBM. In all cases, the prediction target is the formation energy per atom (eV/atom). Each dataset is organized into an ordered domain sequence to evaluate sequential adaptation to incoming data while retaining predictive knowledge acquired from previously observed domains.

Table 2: Statistics of the processed benchmarks. Atom counts refer to the stored input cells. Formation-energy statistics are in eV/atom and summarize all retained records across train, validation, and test partitions; standard deviations use the population convention. Structural groups follow the dataset-specific matching rules.

Statistic	MP	JARVIS-DFT	WBM
Records	132,752	93,902	236,624
Training records	106,208	75,174	189,183
Validation records	13,272	9,309	23,614
Test records	13,272	9,419	23,827
Domains	7	3	5
Structural groups	132,728	93,902	236,624
Atoms/cell: mean	29.15	10.38	7.76
Atoms/cell: range	1–444	1–140	1–100
Energy: mean $\pm$ std.	-1.408 ± 1.163	-0.833 ± 1.047	-0.769 ± 1.117
Energy: range	$[-4.612, 2.500]$	$[-4.423, 5.360]$	$[-10.000, 87.136]$

B.1.1 Materials Project (MP)

Dataset Introduction. The Materials Project (MP) (Jain et al., 2013) is a large-scale open database of inorganic crystalline materials, with properties computed primarily through high-throughput density functional theory (DFT) calculations. It provides relaxed crystal structures together with a broad range of computed material properties and has been widely adopted for training and evaluating machine-learning models for crystalline materials. We use the 132,752 records in `matbench_mp_e_form`, with `e_form` as the prediction target, and retain all records in the prepared benchmark.

Continual-Learning Setup. To construct the continual-learning sequence, we partition the crystals according to their crystallographic systems. Specifically, the seven crystal-system domains are presented sequentially as

Crystal-System Domain Sequence: triclinic $\rightarrow$ monoclinic $\rightarrow$ orthorhombic $\rightarrow$ tetragonal $\rightarrow$ trigonal $\rightarrow$ hexagonal $\rightarrow$ cubic

This construction induces a crystallographic domain shift, where the model sequentially encounters structures belonging to different symmetry families while being expected to preserve predictive knowledge acquired from previously observed crystal systems. The prescribed sequence defines the continual-learning stream and is not intended to represent a strictly increasing scalar measure of structural symmetry.

Data Processing. Each crystal is assigned to exactly one domain according to its crystallographic system. Domain sizes follow the original crystallographic composition of the dataset, without subsampling to equalize the number of samples across domains. Thus, the resulting benchmark preserves the natural imbalance among crystal systems while providing a controlled sequence of crystallographic distribution shifts.

B.1.2 JARVIS-DFT

Dataset Introduction. JARVIS-DFT (Choudhary et al., 2020) is part of the NIST Joint Automated Repository for Various Integrated Simulations (JARVIS) and provides DFT-computed crystal structures and material properties for a large collection of crystalline materials. The database contains quantities such as formation energies, band gaps, elastic properties, dielectric properties, and magnetic properties, and has been continuously expanded as additional materials and calculations become available.

Continual-Learning Setup. Unlike MP, where domains are defined according to crystallographic categories, we use the evolution of the JARVIS-DFT database itself to construct a snapshot-based

continual-learning scenario. Specifically, we use the snapshots dated 2021-08-18, 2022-12-12, and 2025-09-24 and present the resulting domains in chronological order:

Snapshot-Based Domain Sequence: $2021 \rightarrow 2022 \rightarrow 2025$

These dates identify the source snapshots and do not represent the discovery or first-publication time of individual materials. This construction yields a snapshot-based database-expansion scenario, where successive stages introduce materials newly admitted from later releases while previously observed materials remain part of the accumulated knowledge.

Data Processing. Because later releases contain materials that already appear in earlier snapshots, directly treating each snapshot as an independent domain would repeatedly expose the model to the same crystals and blur the distinction between historical and newly arriving data. We therefore assign each material to the first selected snapshot in which its identifier has a valid structure and an unambiguous finite formation-energy label.

Formally, let $\mathcal{I}_t$ denote the set of identifiers with valid and unambiguous records in selected snapshot t. The identifiers first admitted at stage t are defined as

$$\mathcal{N}_t = \mathcal{I}_t \setminus \bigcup_{\tau<t} \mathcal{I}_\tau, \tag{9}$$

and domain $\mathcal{D}_t$ contains the structure–label pairs associated with $\mathcal{N}_t$ in snapshot t. These pairs remain fixed after admission.

Repeated entries with the same identifier and structure are consolidated when a unique finite label is available. If conflicting finite labels occur, the corresponding identifier is deferred until a later snapshot resolves the ambiguity. Once assigned, the same identifier is not introduced again in subsequent domains. The resulting domains contain 55,711, 20,282, and 17,909 records, respectively, yielding a total of 93,902 distinct JARVIS identifiers. The second domain additionally contains one identifier deferred from the first snapshot. Distinct identifiers may still correspond to structurally equivalent crystal inputs.

B.1.3 WBM

Dataset Introduction. The Wang–Botti–Marques (WBM) dataset (Wang et al., 2021) was constructed for exploring previously unknown crystalline materials through chemically informed elemental substitution. Starting from known crystal structures, elements are replaced by chemically similar alternatives according to substitution tendencies derived from existing material data. The generated candidate structures are subsequently evaluated using DFT calculations, providing relaxed crystal structures together with corresponding formation energies.

Continual-Learning Setup. An important property of WBM for our setting is that the generated candidates are organized according to successive chemical-substitution rounds. We use the five official substitution-step files from Materials Cloud, which contain 257,489 structure records before deduplication. We preserve their original round order and treat the five substitution rounds as five consecutive continual-learning domains:

Substitution-Round Domain Sequence: $\mathcal{D}_1 \rightarrow \mathcal{D}_2 \rightarrow \mathcal{D}_3 \rightarrow \mathcal{D}_4 \rightarrow \mathcal{D}_5$

This construction induces a chemically driven distribution shift, where each successive domain corresponds to a different stage of the candidate-generation process. Unlike the crystallographic partition used for MP or the snapshot-based expansion used for JARVIS-DFT, the WBM sequence captures distribution evolution arising from successive chemical-substitution operations. A monotonic increase in distribution distance or prediction difficulty across substitution rounds is not assumed.

Data Processing. The original collection contains multiple records that may correspond to structurally equivalent crystals appearing in different substitution rounds. Without additional processing,

such repeated structures could be counted as new arrivals even when equivalent inputs had already appeared in earlier domains. We therefore perform structural deduplication before constructing the final continual-learning benchmark.

Within each reduced composition, periodic structures are compared using `StructureMatcher`. We use `ltol`=0.001, `stol`=0.001, and `angle_tol`=0.1 degrees, with primitive-cell reduction enabled, lattice rescaling disabled, and supercell and subset matching disabled. Connected components under the resulting matching relation define duplicate groups. For each group, only the earliest occurrence in the substitution sequence is retained, with ties resolved according to the original entry position. Representative selection does not use target values.

This procedure removes 20,865 repeated occurrences while preserving the first-occurrence ordering under the specified structural matching rule. After deduplication, the WBM benchmark contains 236,624 records, with domain sizes of 61,120, 50,278, 71,522, 36,703, and 17,001, respectively. We use the official relaxed crystal structures as inputs and the corresponding `summary.txt` formation energies, unchanged, as prediction targets. Entry alignment is checked against the original sequence, composition, and atom count, while six zero-atom summary placeholders are excluded from this alignment. We apply no additional energy corrections, stability filtering, prototype filtering, or target-value filtering. Accordingly, the observed formation-energy range of $[-10.000, 87.136]$ eV/atom is retained in the benchmark.

B.2 Baseline Details

We provide additional descriptions of the representative continual-learning baselines used in our experiments. For methods originally developed for classification or other graph-learning settings, we retain their core continual-learning mechanisms while adapting the prediction objective to crystal property regression.

General Continual Learning Baselines. **ER** (Rolnick et al., 2019) maintains a memory of historical samples and randomly reuses them together with current-domain data. **SI** (Zenke et al., 2017) accumulates an online estimate of each parameter's contribution to loss reduction along the training trajectory and penalizes changes to important parameters. **MAS** (Aljundi et al., 2018) estimates parameter importance from the sensitivity of model outputs and regularizes important parameters during subsequent learning. **Online EWC** (Schwarz et al., 2018) accumulates parameter importance across domains using an online approximation of the Fisher information. **A-GEM** (Chaudhry et al., 2019) uses gradients from episodic memory as a reference and constrains updates that interfere with previous knowledge. **MIR** (Aljundi et al., 2019) improves replay by prioritizing historical samples predicted to suffer the largest interference from the current update.

Recent Replay and Optimization Baselines. **LPR** (Yoo et al., 2024) stabilizes replay-based continual learning through layer-wise proximal updates that restrict abrupt changes in historical representations. **OCAR** (Urettini & Carta, 2025) incorporates second-order curvature information into replay optimization to stabilize historical knowledge while adapting to new data.

Continual Graph Learning Baselines. **RL-GNN** (Song et al., 2025) learns invariant graph rationales across sequential tasks to reduce forgetting caused by task-specific spurious correlations. **CAGNN** (Xiao et al., 2026) constructs task-specific subnetworks through context-aware hierarchical masking, enabling parameter isolation and knowledge reuse across sequential graph-learning tasks. **HPNs** (Zhang et al., 2023) preserve prior knowledge through hierarchical prototypes encoding node attributes and graph topology, selectively updating relevant prototypes and feature extractors without replaying historical graph data. **SAOT** (Zhang et al., 2026b) employs structure-aware optimal transport together with cross-task knowledge transfer to preserve relational graph structure during continual learning. **TRACE** (Peng et al., 2025) is a self-supervised continual graph learning method inspired by complementary learning systems, in which a fast-learning system extracts key knowledge through node proxies and a slow-learning system consolidates memory via adaptive spaced replay. These graph-specific baselines complement conventional continual-learning methods by explicitly considering structural information in sequential graph data.

B.3 Evaluation Metrics

Consider a sequence of T domains. The held-out test set of domain d is $\mathcal{D}_d^{\text{test}} = \{(X_{d,i}, y_{d,i})\}_{i=1}^{n_d}$, where $X_{d,i}$ is a crystal structure, $y_{d,i}$ is its scalar target property, and n_d is the number of test samples. Let f_t denote the predictor obtained after learning domain t. At each evaluation stage, the same predictor is evaluated on the test sets of all domains encountered so far, without updating its parameters.

Domain-wise prediction errors. For domain $d \leq t$, we compute the mean absolute error $e_{t,d}$ and root mean squared error $r_{t,d}$ as

$$e_{t,d} = \frac{1}{n_d} \sum_{i=1}^{n_d} \left| f_t(X_{d,i}) - y_{d,i} \right|, \tag{10}$$

$$r_{t,d} = \sqrt{\frac{1}{n_d} \sum_{i=1}^{n_d} \left(f_t(X_{d,i}) - y_{d,i} \right)^2}. \tag{11}$$

MAE measures the average magnitude of prediction errors, whereas RMSE gives greater weight to large errors.

Final predictive accuracy. After training on the complete domain sequence, we evaluate f_T on every domain and report

$$\text{MAE}_{\text{final}} = \frac{1}{T} \sum_{d=1}^{T} e_{T,d}, \tag{12}$$

$$\text{RMSE}_{\text{final}} = \frac{1}{T} \sum_{d=1}^{T} r_{T,d}. \tag{13}$$

Both metrics assign equal weight to each domain, irrespective of its test-set size. In particular, final RMSE is computed by first taking the square root within each domain and then averaging the resulting domain-wise RMSE values. It therefore differs from computing a single RMSE over the pooled test samples. These metrics characterize the final predictor's performance across all learned domains, including the most recent domain.

Forgetting. For each historical domain $d < T$, its reference error is the lowest test MAE attained from its introduction through the stage immediately preceding the final stage:

$$b_d = \min_{d \leq t < T} e_{t,d}. \tag{14}$$

We measure final forgetting as

$$F_{\text{final}} = \frac{1}{T-1} \sum_{d=1}^{T-1} \max\left(0, e_{T,d} - b_d\right). \tag{15}$$

The nonnegative truncation ensures that improvements on one domain do not offset degradation on another. The last domain is excluded because no subsequent domain is learned. The historical minimum is used for retrospective metric calculation and does not define a checkpoint-selection rule.

Final forgetting measures an absolute increase in MAE, without normalization by the reference error. All three metrics are therefore expressed in eV/atom, with lower values indicating better performance. A forgetting value of zero means that the final error does not exceed the historical reference; it does not imply accurate predictions. We consequently interpret forgetting jointly with final MAE and RMSE.

C Additional Experimental Results

C.1 Knowledge Retention under Continual Crystal Evolution

As shown in Figure 8, SGCR exhibits low forgetting across the three datasets, indicating good knowledge retention under evolving crystal distributions. On MP and JARVIS, SGCR achieves the lowest forgetting among the compared methods, whereas several conventional continual-learning baselines show noticeably larger degradation. On WBM, SGCR remains competitive and exhibits lower forgetting than most replay-based and graph continual-learning approaches. Competing methods vary more across datasets, whereas SGCR maintains relatively low forgetting under the different distribution shifts. These observations suggest that combining structure-aware replay with adaptive gradient regulation helps mitigate catastrophic forgetting while accommodating newly emerging crystal domains.

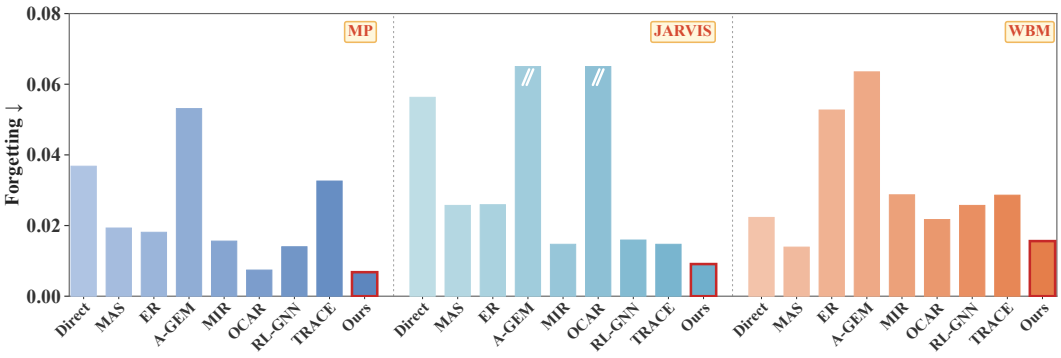

Figure 8: Forgetting comparison using CGCNN as the backbone.

C.2 Main Results with Matformer as Backbone

Table 3: Performance comparison on MP, JARVIS, and WBM with Matformer as backbone. MAE and RMSE are final domain-macro averages in eV/atom. For each baseline, the orange percentage indicates its relative error increase over SGCR. The **best** results and second-best results are highlighted.

Type	Method	MP		JARVIS		WBM		Average	
		MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
General CL	Direct	0.1031 ↑26.04%	0.1443 ↑19.35%	0.1137 ↑35.04%	0.1920 ↑42.43%	0.1351 ↑9.93%	0.7413 ↑0.26%	0.1173 ↑21.81%	0.3592 ↑8.29%
	SI [ICML 2017]	0.1576 ↑92.67%	0.2296 ↑89.91%	0.1012 ↑20.19%	0.1426 ↑5.79%	0.1667 ↑35.64%	0.7500 ↑1.43%	0.1418 ↑47.25%	0.3741 ↑12.78%
	MAS [ECCV 2018]	0.1646 ↑101.22%	0.2348 ↑94.21%	0.1108 ↑31.59%	0.2076 ↑54.01%	0.1610 ↑31.00%	0.7473 ↑1.07%	0.1455 ↑51.09%	0.3966 ↑19.57%
	ER [NeurIPS 2019]	0.0845 ↑3.30%	0.1223 ↑1.16%	0.0856 ↑1.66%	0.1438 ↑6.68%	0.1310 ↑6.59%	0.7685 ↑3.94%	0.1004 ↑4.26%	0.3449 ↑3.98%
	A-GEM [ICLR 2019]	0.1185 ↑44.87%	0.1793 ↑48.30%	0.0974 ↑15.68%	0.1580 ↑17.21%	0.1268 ↑3.17%	0.7424 ↑0.41%	0.1142 ↑18.59%	0.3599 ↑8.50%
	MIR [NeurIPS 2019]	0.1214 ↑48.41%	0.1842 ↑52.36%	0.1044 ↑23.99%	0.1730 ↑28.34%	0.1367 ↑11.23%	0.7614 ↑2.98%	0.1208 ↑25.44%	0.3729 ↑12.42%
	Online EWC [ICML 2018]	0.1368 ↑67.24%	0.1931 ↑59.72%	0.1067 ↑26.72%	0.1737 ↑28.86%	0.1261 ↑2.60%	0.7545 ↑2.04%	0.1232 ↑27.93%	0.3738 ↑12.69%
	OCAR [ICML 2025]	0.1157 ↑41.44%	0.1782 ↑47.39%	0.0897 ↑6.53%	0.1510 ↑12.02%	0.1345 ↑9.44%	0.7582 ↑2.54%	0.1133 ↑17.65%	0.3625 ↑9.29%
	LPR [ICML 2024]	0.1668 ↑103.91%	0.2248 ↑85.94%	0.0897 ↑6.53%	0.1617 ↑19.96%	0.1371 ↑11.55%	0.7640 ↑3.33%	0.1312 ↑36.24%	0.3835 ↑15.62%
Graph CL	HPNs [TPAMI 2023]	0.1171 ↑43.15%	0.1570 ↑29.86%	0.0873 ↑3.68%	0.1396 ↑3.56%	0.1257 ↑2.28%	0.7428 ↑0.46%	0.1100 ↑14.23%	0.3465 ↑4.46%
	CAGNN [ICML 2026]	0.1178 ↑44.01%	0.1790 ↑48.06%	0.0943 ↑12.00%	0.1407 ↑4.38%	0.1645 ↑33.85%	0.7768 ↑5.06%	0.1256 ↑30.43%	0.3655 ↑10.19%
	SAOT [ICML 2026]	0.1674 ↑104.65%	0.2333 ↑92.97%	0.2987 ↑254.75%	0.4046 ↑200.15%	0.2623 ↑113.43%	0.8416 ↑13.82%	0.2428 ↑152.13%	0.4931 ↑48.66%
	RL-GNN [AAAI 2025]	0.0868 ↑6.11%	0.1248 ↑3.23%	0.0938 ↑11.40%	0.1476 ↑9.50%	0.1424 ↑15.87%	0.7516 ↑1.65%	0.1077 ↑11.84%	0.3413 ↑2.89%
SGCR (Ours)		**0.0818**	**0.1209**	**0.0842**	**0.1348**	**0.1229**	**0.7394**	**0.0963**	**0.3317**

To further evaluate the generalizability of different continual learning methods across backbone architectures, we conduct additional experiments using Matformer as the backbone, with the results reported in Table 3. With Matformer, the baseline methods exhibit clear dataset-dependent behavior. Among the general continual learning methods, ER is particularly competitive, achieving the strongest baseline results on MP and JARVIS in MAE, as well as the lowest average MAE. Among graph-specific methods, RL-GNN achieves the lowest baseline average RMSE. Nevertheless, no single baseline consistently dominates across all datasets and metrics, indicating that their effectiveness depends on the underlying domain shift. SGCR achieves the lowest mean error across all reported

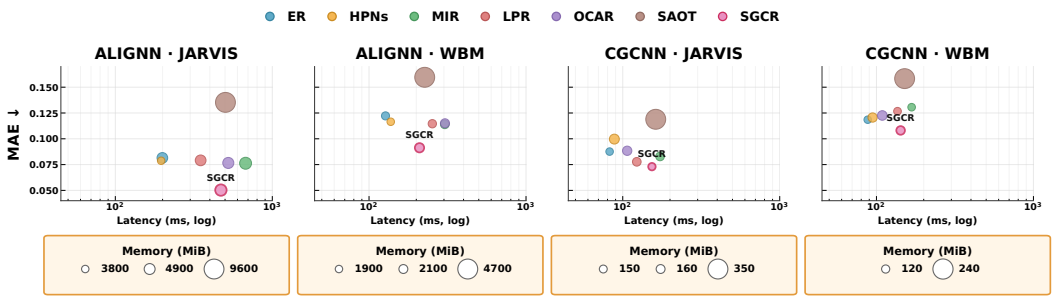

Figure 9: **Accuracy–efficiency trade-off on JARVIS and WBM with ALIGNN and CGCNN.** The x-axis denotes per-step training latency (log scale), the y-axis denotes final MAE, and bubble size represents peak training memory. Bubble sizes are normalized independently within each panel to improve visibility, with the corresponding memory scales reported below each subplot. Lower MAE, lower latency, and smaller memory consumption are preferable.

metrics, although its margins over the strongest baseline are small in several cases, e.g., 0.26% for WBM RMSE and 1.14% for MP RMSE, and run-to-run variance is not reported for this backbone.

C.3 Computational Efficiency Analysis

We further examine the computational efficiency of SGCR by jointly considering predictive accuracy, training latency, and peak memory consumption. Specifically, we measure the per-step training latency and peak GPU memory at the final continual-learning stage, i.e., D3 for JARVIS and D5 for WBM. As shown in Figure 9, the horizontal axis represents training latency on a logarithmic scale, the vertical axis reports final MAE, and the bubble size indicates peak memory usage. Since ALIGNN and CGCNN exhibit substantially different memory footprints, bubble sizes are normalized independently within each panel, with the corresponding memory scales shown below each subplot.

SGCR achieves the lowest MAE in all four backbone–dataset settings while maintaining a competitive computational footprint. With ALIGNN, SGCR requires 470.49 and 209.65 ms/step on JARVIS and WBM, respectively, which are approximately 30.4% and 31.0% faster than MIR. Compared with SAOT, SGCR reduces peak memory by 47.0% on JARVIS and 53.8% on WBM, while also achieving lower latency. Similar behavior is observed with CGCNN: SGCR is 11.2% and 14.9% faster than MIR on JARVIS and WBM, respectively, and requires only 146.16 and 122.86 MiB of peak memory. In particular, its memory footprint on CGCNN–JARVIS is essentially identical to that of the lightweight ER baseline (146.16 vs. 146.25 MiB).

Compared with simpler replay strategies such as ER, SGCR incurs additional per-step computation due to structure-aware retrieval, conflict-aware replay, and gradient correction. Nevertheless, the additional cost is accompanied by lower prediction error without excessive memory growth. Overall, these results indicate that SGCR provides a favorable balance between predictive accuracy and computational efficiency across different graph backbones and continual-learning streams.

C.4 Discussion and Future Directions

While our experiments demonstrate the effectiveness of SGCR across different crystal-evolution scenarios and graph backbones, several directions remain worth exploring.

Broader material properties. The present study focuses on continual formation-energy prediction with a scalar regression objective. Future work could extend SGCR to other material properties and multi-property prediction tasks, further examining its applicability across diverse materials-discovery scenarios.

Large-scale continual material streams. As materials databases continue to grow, more efficient implementations of structural indexing and replay retrieval could be explored, such as compressed

prototype representations or approximate search strategies. These extensions may further improve the scalability of SGCR to substantially larger continual material streams.

Domain-free continual learning. Our current experiments consider explicitly defined sequential domains. Future work could investigate domain-free settings, where distribution changes are identified implicitly from incoming crystal data and replay allocation is dynamically adjusted according to the evolving structural distribution.