Instituto Tecnológico y de Estudios Superiores de Monterrey

TC5035.10 – Proyecto Integrador

Grupo 10

Avance 3. Baseline

Equipo 17

Alumno	Matrícula
Carlos Giovanny Encinia González	A01795601
Gustavo Pérez Juárez	A01795310
Ignacio Antonio Quintero Chávez	A01794419

Profesor Titular: Dra. Grettel Barceló Alonso

Profesor Sponsor: Dr. Juan Arturo Nolazco-Flores

12 de octubre de 2025

Tabla de contenidos

Introducción

En este documento tenemos como objetivo evaluar la capacidad del modelo generativo MatterGen, desarrollado por Microsoft Research, para sintetizar materiales más allá de su dominio de entrenamiento original.

En particular, buscamos determinar si es posible utilizar MatterGen para generar o aproximar las siguientes clases de materiales:

1) Materiales bidimensionales (2D),
2) Grafeno epitaxial intercalado con oro (Au–C),
3) Estructuras Metal–Orgánicas (MOFs),
4) Materiales semiconductores derivados del grafeno.

El modelo MatterGen fue entrenado principalmente con materiales individuales provenientes del *Materials Project Database*, que incluye óxidos, sulfuros y semiconductores convencionales, por lo que no fue diseñado originalmente para producir combinaciones arbitrarias de elementos (por ejemplo, carbono con metales nobles).

El propósito de este avance es establecer una línea base (*baseline*) que permita analizar la viabilidad de utilizar MatterGen como herramienta exploratoria para la generación de nuevos materiales híbridos o no incluidos en su conjunto de entrenamiento.

Selección del algoritmo baseline y justificación

Aunque MatterGen se basa en una arquitectura de modelos de difusión generativos condicionados, el *baseline* aquí se define en términos de su capacidad para generalizar fuera de su dominio de entrenamiento.

Se emplea el modelo preentrenado *dft_band_gap*, el cual genera materiales condicionados a una propiedad electrónica fundamental: el *band gap* calculado mediante teoría del funcional de la densidad (DFT).

El *band gap* permite clasificar materiales según su comportamiento electrónico:

1) Metales y semimetales: *band gap* ≈ 0.0 eV
2) Semiconductores: *band gap* entre 0.5–3.0 eV
3) Aislantes: *band gap* > 3.0 eV

Condicionar el modelo a esta propiedad posibilita evaluar si puede reproducir estructuras que correspondan a materiales metálicos, semiconductores o aislantes, aun sin especificar directamente su composición química. Por lo tanto, este *baseline* se define a continuación.

Evaluar si MatterGen puede generar estructuras físicas y químicamente coherentes al condicionarse únicamente a propiedades electrónicas (*band gap*), sin requerir combinaciones explícitas de elementos.

Análisis de características

Debido a que el modelo no permite condicionar directamente sobre la composición química (*chemical_system*), se recurrió a propiedades físicas indirectas que reflejan el tipo de material buscado.

Estas propiedades se utilizaron como parámetros de entrada del modelo y como criterios de filtrado de los resultados generados.

Propiedad	**Rango objetivo**	**Tipo de material asociado**
dft_band_gap ≈ 0.0 ± 0.1 eV	Metales / semimetales	Grafeno, Au–C
dft_band_gap ≈ 1.8 ± 0.3 eV	Semiconductores	Grafeno dopado
density < 3.0 g/cm³	Materiales porosos	Estructuras metal-orgánicas (MOFs)
formation_energy_per_atom < 0	Estabilidad química	Validación estructural

Estas propiedades fueron seleccionadas por su relevancia en la caracterización de materiales cristalinos y por ser *features* que MatterGen utiliza internamente para la generación condicionada.

Por ejemplo, al condicionar el modelo a un *band gap* de 0.0 eV, se obtuvo la generación de estructuras con carácter metálico o semimetálico, algunas de las cuales contenían carbono (C) y oro (Au), coincidiendo con el comportamiento esperado de materiales conductores.

Evaluación del modelo

Subajuste y sobreajuste

MatterGen es un modelo preentrenado, por lo que no se realizaron ajustes ni reentrenamientos sobre los datos originales. No obstante, desde un punto de vista conceptual, se puede hablar de un subajuste fuera de dominio, ya que el modelo:

1) No logra representar correctamente combinaciones de elementos no vistas durante su entrenamiento (por ejemplo, Au–C),
2) Tiende a reproducir materiales similares a los de su dataset original (óxidos, sulfuros o semiconductores binarios).

Esto refleja una limitación de generalización más que un problema de sobreajuste clásico, evidenciando que MatterGen opera de manera confiable dentro de su dominio pero presenta dificultades al extrapolar hacia nuevos sistemas químicos.

Desempeño mínimo aceptable

Dado que MatterGen no fue entrenado con materiales de tipo híbrido o bidimensional, se propone un umbral de rendimiento mínimo aceptable del 20 % para la tasa de estructuras válidas (TEV). Esto implica que, si al menos una de cada cinco estructuras generadas cumple con los criterios de validez física y composicional, el modelo puede considerarse útil como herramienta exploratoria fuera de dominio.

Este umbral se considera razonable para un modelo no ajustado y permite establecer una referencia objetiva para futuras mejoras o experimentos de *fine-tuning* con *datasets* especializados.

Conclusiones

Los resultados obtenidos permiten establecer una línea base conceptual sobre el desempeño de MatterGen fuera de su dominio de entrenamiento.

El modelo muestra una alta coherencia interna al generar materiales dentro de los rangos esperados de *band gap*, densidad y estabilidad; sin embargo, su capacidad de extrapolación hacia combinaciones químicas nuevas es limitada.

A pesar de ello, la generación condicionada por propiedades físicas demuestra ser una estrategia viable para inducir indirectamente la creación de materiales con comportamientos electrónicos específicos, como los grafenos dopados o semimetales.

En consecuencia, MatterGen puede considerarse una herramienta prometedora para la exploración guiada de materiales complejos, particularmente si se complementa con procesos de *fine-tuning* o bases de datos extendidas que incluyan materiales bidimensionales, MOFs y heteroestructuras metal–carbono.

Referencias

Microsoft Research. (2025). MatterGen: A new paradigm of materials design with generative AI. Recuperado de: https://www.microsoft.com/enus/research/blog/mattergen-a-new-paradigm-of-materials-design-with-generative-ai/.

ScienceDirect. (2024). A guide to discovering next-generation semiconductor materials using atomistic simulations and machine learning. Current Opinion in Solid State and Materials Science, 28(3), 101148. Recuperado de: https://www.sciencedirect.com/science/article/abs/pii/S092702562400329X