Este trabajo construye una primera ontología operativa para una arquitectura de IA aplicada al diseño y estudio de superconductores. En la visión planteada en el documento de avances, la ontología organiza el proceso de síntesis en ocho etapas y aquí se diseña la primera etapada ("Diseño molecular y composicional") mediante modelos de *machine learning* que predicen la temperatura crítica T_c a partir de descriptores fisicoquímicos.

Se utilizó MDR SuperCon Datasheet, repositorio público mantenido por NIMS, que publica versiones con DOI y organiza la información en tablas separadas para **"**Oxide & Metallic" y "Organic", además de un *preview* "primary.tsv" pensado para *machine learning*. El esquema incluye fórmulas químicas, composiciones elementales y un amplio conjunto de propiedades/mediciones, entre ellas T_cy su método de medida.

En particular, las columnas Tc recomendada (tc), año de publicación (year) y método de medición de Tc (tcmeth) forman parte del diccionario del dataset y permiten filtrar, normalizar y auditar coherencia.

Preprocesamiento

1. **Limpieza del objetivo** T_c: exclusión de registros sin T_co con inconsistencias básicas.
2. **Control de calidad**: deduplicación por material-publicación y banderas simples (unidad, rangos plausibles).
3. **Normalización**: estandarización de tipos y manejo de faltantes.
4. **Ingeniería de características**:
    - **Bag-of-elements** (fracciones atómicas por elemento).
    - **Descriptores "Magpie"**: estadísticos sobre propiedades elementales (media ponderada, rango, desviación, entropía, etc.), siguiendo la formulación habitual para propiedades $P_k$ponderadas por fracciones c_k. Estas familias incluyen, entre otras, masas atómicas, radios, electronegatividad y conductividad térmica.

Construcción de vistas del problema

- **Vista A (nivel material)**: una fila por material, agregando sus mediciones de T_c(mediana e IQR).
- **Vista B (nivel publicación-medición)**: conserva cada medición original y su referencia temporal.

Esta separación permite comparar robustez a ruido (A) frente a capacidad de explotar variabilidad experimental (B).

Se entrenaron Random Forest (RF) y XGBoost (XGB), con validación cruzada GroupKFold y evaluación adicional de validación temporal (entrenar con artículos hasta un año de corte y probar en años posteriores). El manuscrito de avances justifica la elección de árboles de conjunto por su buen rendimiento en tabulares y su capacidad intrínseca de estimar importancias.

Métricas de validación cruzada (promedio CV):

- **Vista A**: RF MAE 4.82 K, RMSE 9.49 K, $R^2$0.914; XGB MAE 5.01 K, RMSE 9.36 K, $R^2$0.917.
- **Vista B**: RF MAE 8.15 K, RMSE 15.76 K, $R^2$0.711; XGB MAE 8.37 K, RMSE 15.64 K, $R^2$0.713.

Validación temporal (cortes 1995, 2005, 2015): el error decrece al acercar el corte al presente y R^2se mantiene moderado (≈0.49–0.71 según modelo y corte), evidenciando desfase de distribución entre épocas y el valor de re-entrenar con literatura reciente.

Resultados principales

1) Desempeño predictivo (regresión)

La **Vista A** supera con claridad a la **Vista B** en MAE/RMSE, lo que sugiere que **agregar por material** amortigua ruido experimental y sesgos de reporte (gráficas de barras MAE/RMSE/R^2, Fig. 1).

2) Explicabilidad por importancias

En ambas familias de modelos, la variable **mag_thermal_conductivity_range** domina la señal (≈0.54 en RF; ≈0.77 en XGB), seguida por proporciones elementales (O, Ba, Ca, Cu) y dispersión en radio/masa atómica.

Este hallazgo es coherente con la discusión del manuscrito, donde se subraya que la **explicabilidad de árbol** revela **asociaciones predictivas** (no causalidad física) y se propone una lectura fisicoquímica de las variables dominantes.

Una alta contribución del rango de conductividad térmica implica que heterogeneidades fuertes entre los elementos del compuesto están estrechamente correlacionadas con variaciones en $T_c$dentro del dominio de datos; no significa que el modelo "comprenda" mecanismos microscópicos (p. ej., espectros fonónicos), sino que esa característica permite particionar el espacio de decisión con bajo error.

En las Fig. 2 (top-5 por modelo) se aprecia la jerarquía de descriptores y refuerza la convergencia de ambos algoritmos en el mismo driver principal.

3) Diagnósticos de ajuste

- **Paridad y residuales (Fig. 3):** la nube se alinea con la diagonal y los residuales se concentran en torno a 0 K, con dispersión mayor en T_c altos (heterocedasticidad esperable).
- **Residual vs. $\hat{T}_c$(Fig. 3, derecha):** se observa abanico para $\hat{T}_c$ intermedios-altos, útil para definir bandas de predicción y priorizar verificación experimental.

4) Incertidumbre (cuantiles)

Se calcularon intervalos $p10$–$p90$ por predicción; los MAE por cuantil fueron 13.86 K (p10), 6.17 K (p50) y 10.99 K (p90) y la cobertura empírica global del 80% objetivo fue ≈0.74, con variaciones por decil de $\hat{T}_c$(Fig. 4).

5) Clasificación de alto T_c(umbral 77 K)

Para apoyar priorización de candidatos, se formuló una tarea binaria $T_c \geq 77$K con XGBoost: ROC-AUC≈0.987, PR-AUC≈0.926 y F1≈0.859 con umbral ≈0.26 (curvas ROC/PR en Fig. 5).

Relación con la ontología propuesta

Los resultados validan el módulo predictivo de la primera etapa y muestran que descriptores composicionales (estadísticos de propiedades elementales y fracciones) contienen suficiente información para alcanzar errores ~5–10 K y excelente discriminación de alto T_c. Esto respalda la idea de integrar el modelo dentro de una arquitectura iterativa donde etapas posteriores (síntesis, caracterización) retroalimenten el *re-training* y la generación de nuevas características.

Cómo interpretar las figuras incluidas

- **Fig. 1 (barras CV)**: sintetiza en una sola vista que Vista A es más estable que Vista B; ideal para la sección de "Selección de representación del dato".
- **Fig. 2 (importancias)**: vincula *drivers* composicionales con hipótesis fisicoquímicas propuestas en el manuscrito (conductividad térmica, electronegatividad, radios/masas).
- **Fig. 3 (paridad/residuales)**: justifica el uso de intervalos predictivos y advierte sobre heterocedasticidad.
- **Fig. 4 (ancho y cobertura $p10$–$p90$)**: comunica incertidumbre operativa por decil de $\hat{T}_c$, útil para gestión de riesgo experimental.
- **Fig. 5 (ROC y PR)**: traduce el modelo en un criterio de filtrado de materiales candidatos con una línea base cuantitativa para *recall/precision*.

Limitaciones y amenazas a la validez

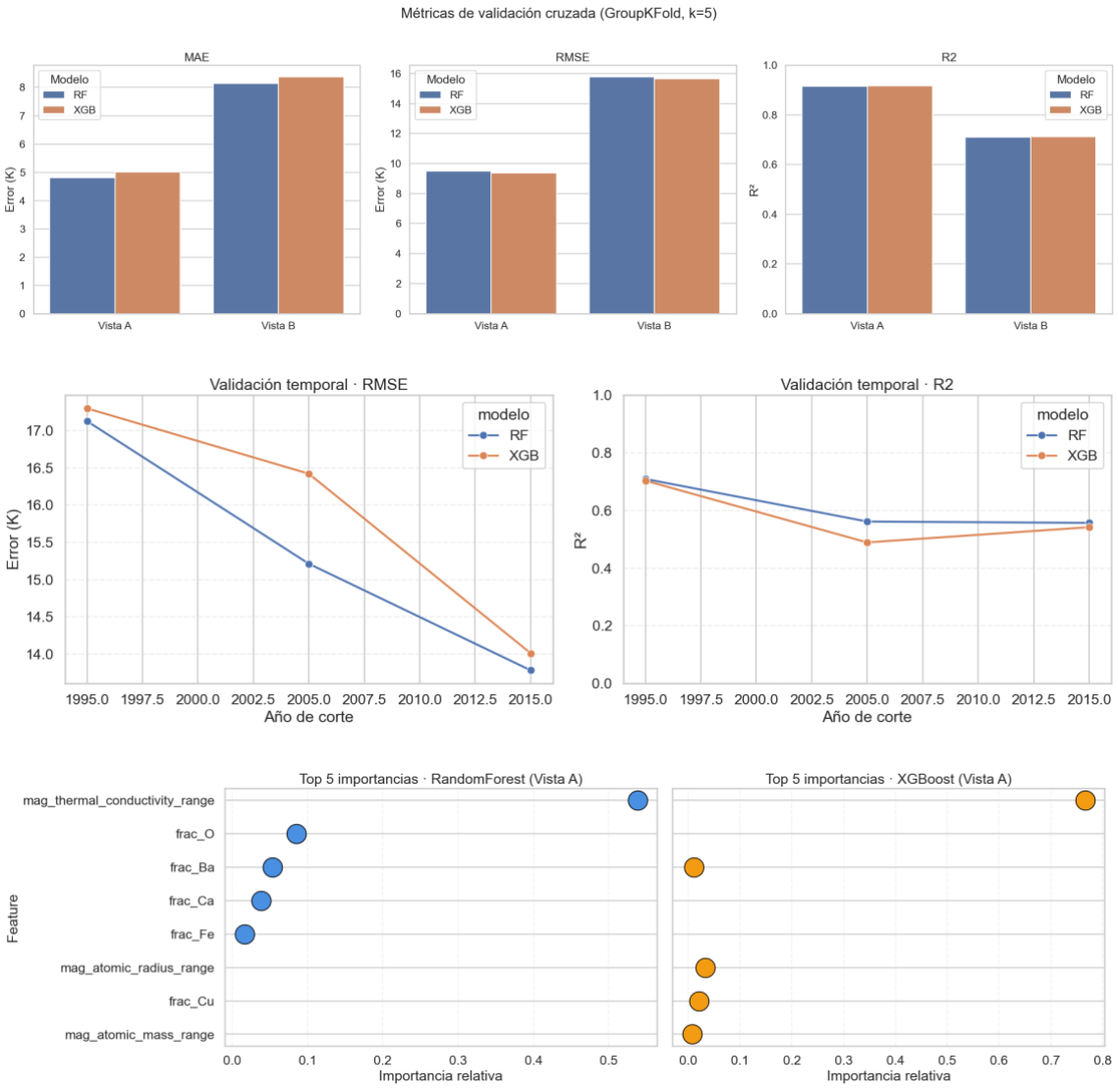
- **No-causalidad**: las importancias son **asociativas**; requieren contraste con teoría/experimentos (el documentos de avances lo discute con claridad).
- **Cambio de distribución temporal**: la degradación de R^2en escenarios de *forecasting* sugiere sesgos de reporte y evolución de familias químicas; mitigable con re-entrenamientos periódicos y *time-aware CV*.

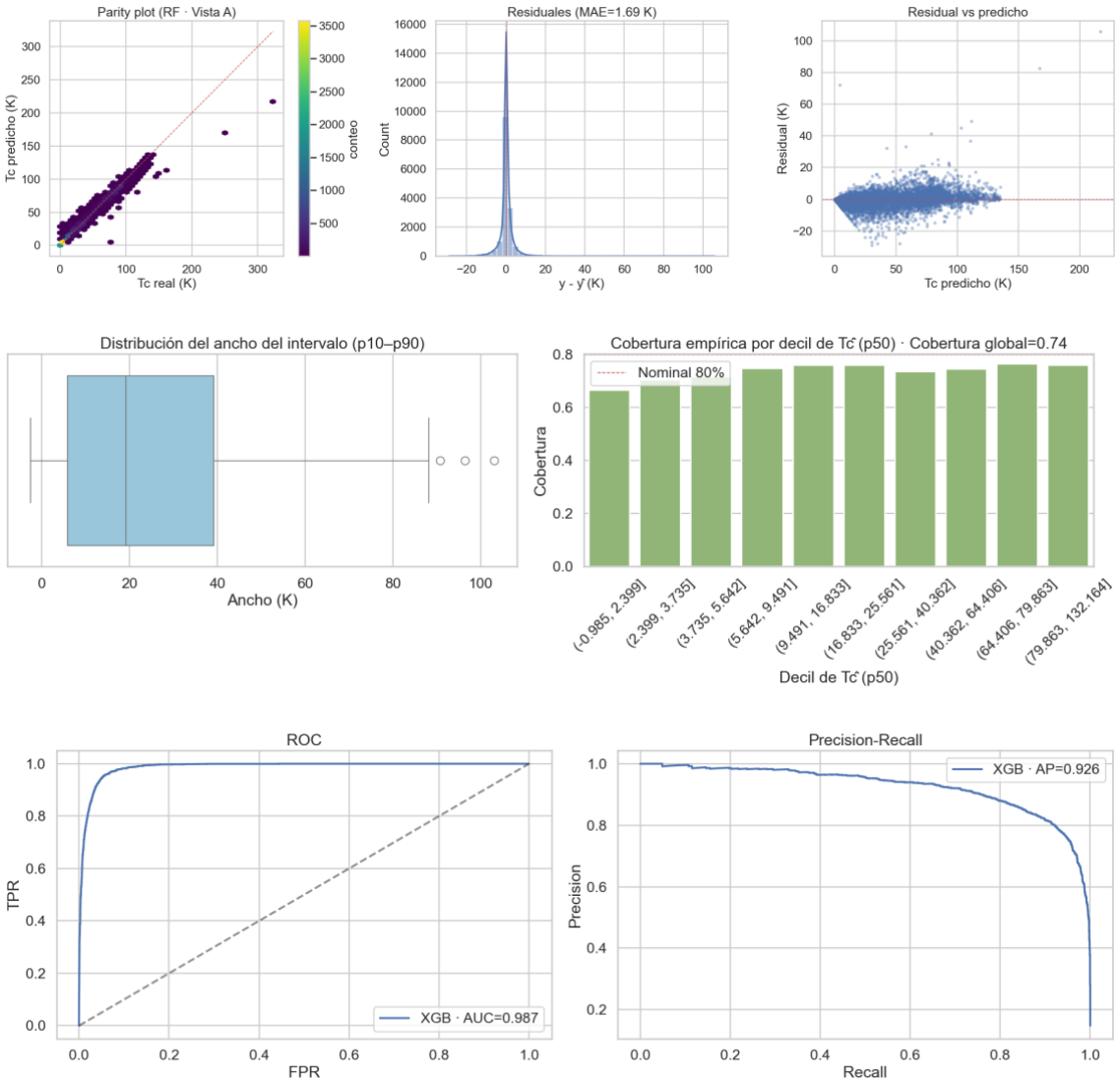
- **Granularidad de SuperCon**: mezcla de **material-publicación** y diversidad de métodos de medida; de ahí la utilidad de la **Vista A** y del campo de método de T_c.

Conclusión operativa

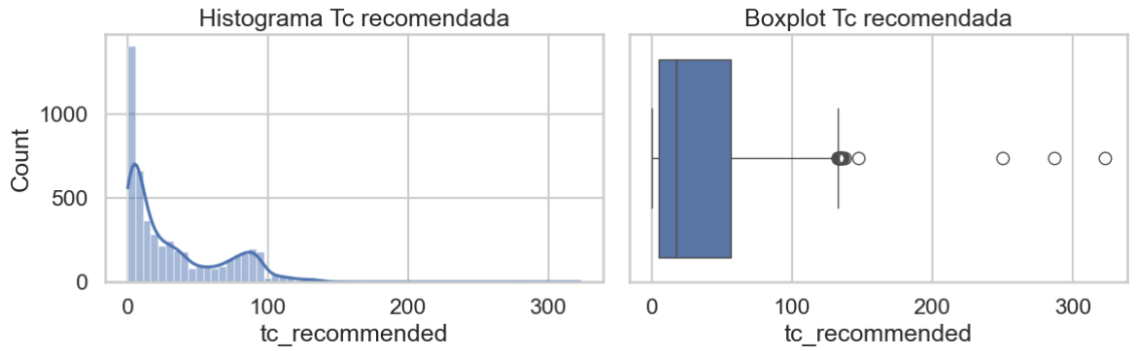
El estudio entrega una **línea base explicable y cuantificada** para T_c:

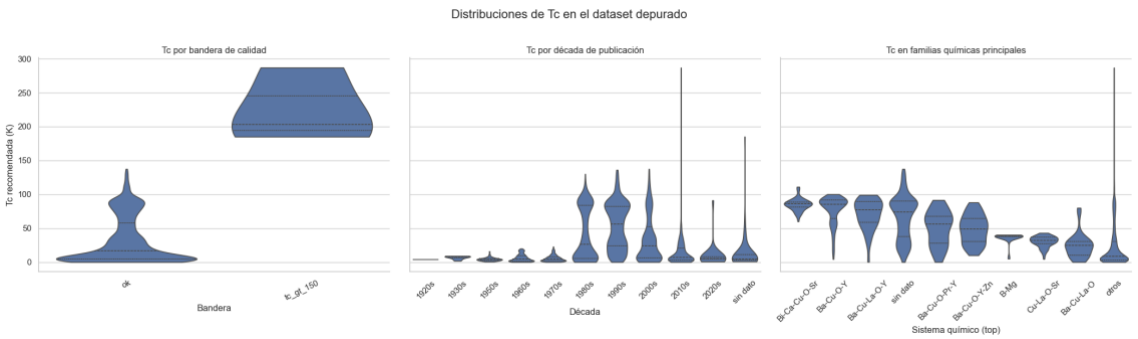
- Predicción en la escala de 5–10 K (Vista A) y priorización de alto $T_c$con AUC~0.99.
- Drivers composicionales consistentes (heterogeneidad en conductividad térmica y estadísticos atómicos) como hipótesis para experimentos dirigidos.
- Plantilla metodológica alineada con la ontología del proyecto para integrarse con etapas posteriores de síntesis y caracterización.





Gráficos de distribución:




Distribuciones de Tc en el dataset depurado
Tc por bandera de calidad
Tc recomendada (K)
300
250
200
150
100
50
0
ok
tc_gt_150
Bandera
Tc por década de publicación
1920s
1930s
1950s
1960s
1970s
1980s
1990s
2000s
2010s
2020s
sin dato
Década
Tc en familias químicas principales
Bi-Ca-Cu-O-Sr
Ba-Cu-O-Y
Ba-Cu-La-O-Y
sin dato
Ba-Cu-O-Pr-Y
Ba-Cu-O-Y-Zn
B-Mg
Cu-La-O-Sr
Ba-Cu-La-O
otros
Sistema químico (top)