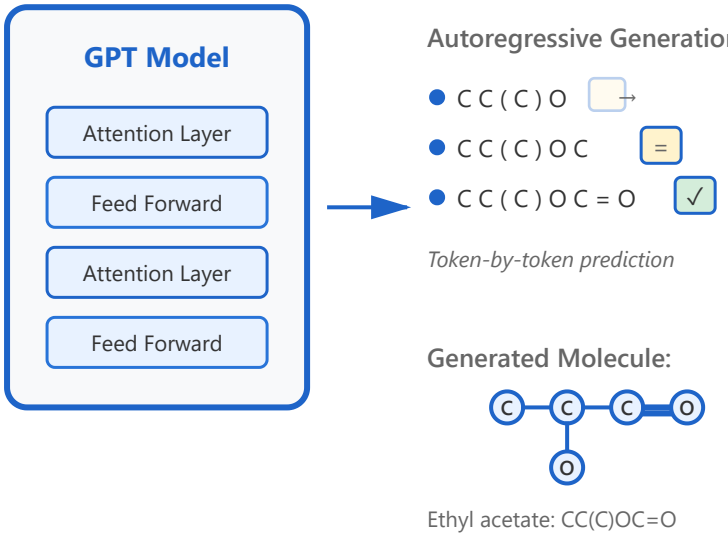
GPT for Molecules

SMILES Generation Process



Ethyl acetate: CC(C)OC=O

SMILES Generation

String-based molecular representation

Property Conditioning

Control molecular attributes

Reaction Prediction

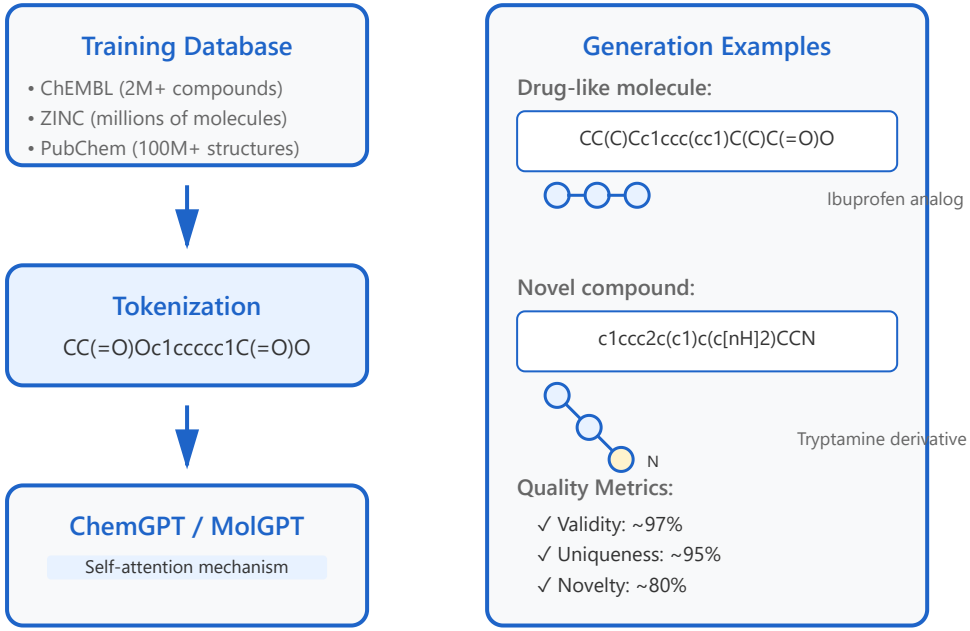
Reactants → Products mapping

Retrosynthesis

Backward synthesis planning

SMILES Generation with Chemical Language Models

Molecular Generation Pipeline



SMILES Representation

SMILES (Simplified Molecular Input Line Entry System) converts molecular structures into sequential text strings, enabling language models to generate valid chemical structures.

Key Features

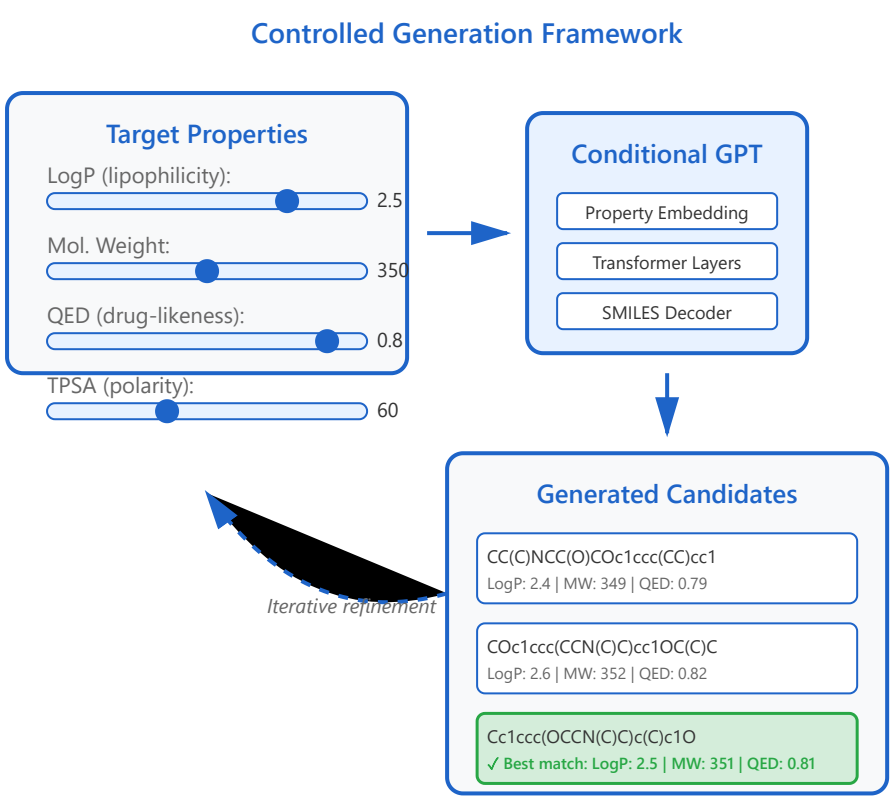
- Character-level or token-level encoding of molecular structure
- Autoregressive generation learns chemical syntax and rules
- Pre-training on massive molecular databases
- Fine-tuning for specific property targets

Applications

- De novo drug design and lead optimization
- Chemical space exploration
- Molecule optimization for desired properties

Models: ChemGPT, MolGPT, and SMILES-BERT leverage transformer architectures to understand chemical grammar and generate chemically valid molecules with high success rates.

Property-Conditioned Molecular Generation



Controlled Generation

Property-conditioned models generate molecules that satisfy specific physicochemical or biological constraints by integrating target properties into the generation process.

Conditioning Approaches

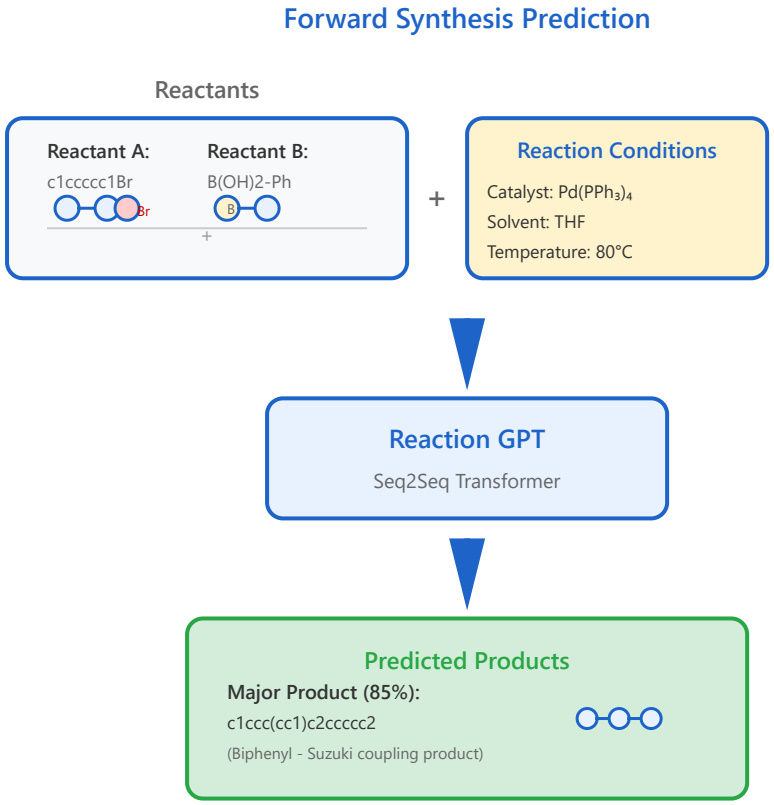
- **Prefix conditioning:** Property tokens prepended to SMILES
- **Latent conditioning:** Property embeddings in hidden layers
- **Reinforcement learning:** Reward-guided optimization
- **Multi-objective:** Balance multiple property constraints

Target Properties

- Physicochemical: LogP, molecular weight, TPSA
- Drug-likeness: QED, Lipinski's rule compliance
- Biological activity: Binding affinity, selectivity
- ADMET: Solubility, permeability, toxicity

Impact: Enables rational drug design by generating molecules with optimized pharmacokinetic properties, reducing experimental screening costs and accelerating lead discovery.

Chemical Reaction Prediction



Forward Reaction Prediction

GPT models learn to predict reaction outcomes by training on millions of reaction examples, mapping reactants and conditions to products using sequence-to-sequence architectures.

Model Architecture

- **Input:** Reactant SMILES + reaction conditions + reagents
- **Encoder:** Processes reactant molecular structure
- **Decoder:** Generates product SMILES sequentially
- **Attention:** Focuses on reactive sites and functional groups

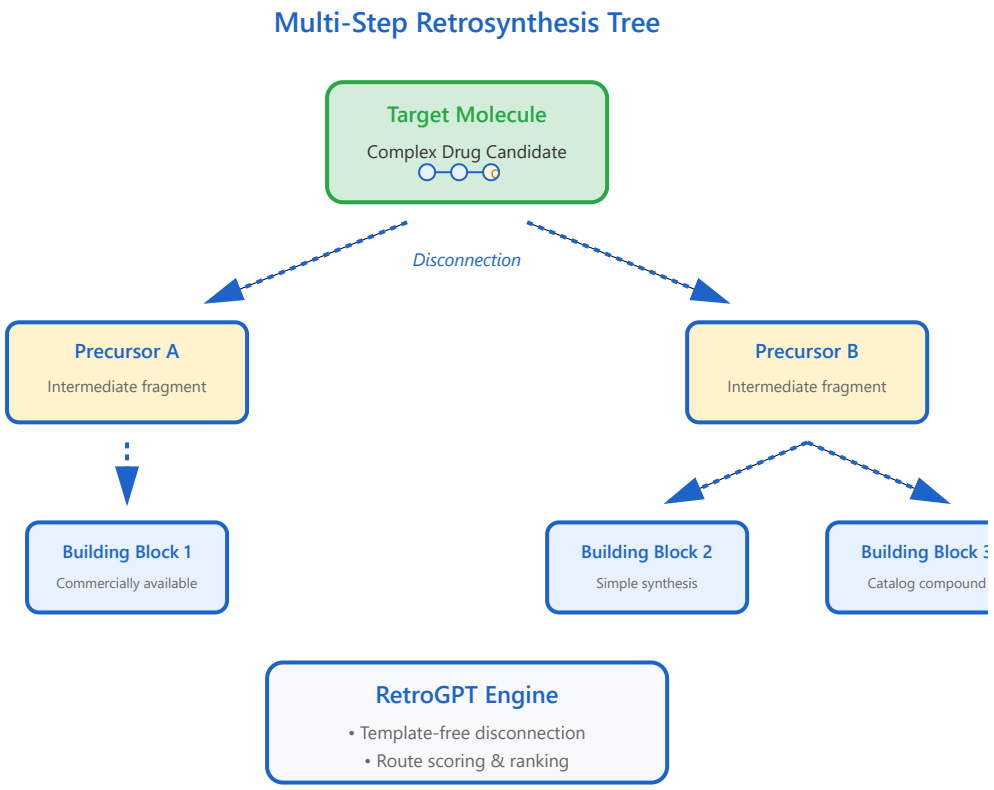
Key Capabilities

- Named reaction prediction (e.g., Suzuki, Grignard, Diels-Alder)
- Regioselectivity and stereochemistry prediction
- Side product and byproduct identification
- Yield estimation and reaction feasibility

Training Data: Models trained on USPTO (United States Patent and Trademark Office) dataset containing millions of experimentally validated

reactions, achieving >90% top-1 accuracy for common reaction types.

Retrosynthetic Planning with GPT Models



Retrosynthetic Analysis

Retrosynthesis works backwards from target molecules to identify synthetic routes using simpler, commercially available starting materials. GPT models automate this complex planning process.

Model Approaches

- **Template-free:** Direct SMILES transformation without predefined reaction rules
- **Template-based:** Apply learned reaction templates from databases
- **Hybrid:** Combine both approaches for robust predictions
- **Multi-step planning:** Build complete synthesis trees

Key Features

- Automated disconnection site identification
- Route feasibility scoring and ranking
- Cost and availability optimization
- Stereoselective synthesis planning

Applications: Retrosynthetic GPT models like Molecular Transformer and RetroGPT are used in pharmaceutical companies to accelerate drug discovery, reducing synthesis planning time from

weeks to hours while suggesting novel synthetic routes.