A well-chosen project in **descriptor calculation using RDKit** can demonstrate:

- computational chemistry skills
- cheminformatics fluency
- coding proficiency (Python, RDKit, Pandas, Matplotlib, etc.)
- scientific thinking (e.g., hypothesis testing, validation)

High-Impact Project Ideas for Your CV

1. QSAR Modeling of Bioactivity Using RDKit Descriptors

> **Goal:** Predict biological activity (e.g., IC_{50}, Ki) from chemical structure using molecular descriptors.

- **Data:** ChEMBL or PubChem bioassays
- **Steps:**
  - Fetch dataset (SMILES + activity)
  - Compute descriptors/fingerprints with RDKit
  - Train ML models (RandomForest, XGBoost, etc.)
  - Validate model (R^2, MAE, ROC-AUC depending on regression/classification)
- **Deliverable:** Jupyter notebook + performance plots + feature importance

Why it's good for your CV: Shows end-to-end pipeline — data wrangling, descriptor generation, modeling, evaluation.

---

2. Descriptor Clustering of Drug-like Compounds

> **Goal:** Use descriptors to group similar molecules and discover clusters of similar bioactivity or scaffolds.

- **Data:** DrugBank, ZINC15, or ChEMBL
- **Techniques:**
  - Generate descriptor matrix (e.g., Morgan fingerprints, MACCS keys)
  - Apply PCA or t-SNE to reduce dimensionality
  - Use DBSCAN or KMeans for clustering
  - Visualize clusters (label top drugs per cluster)

- **Bonus:** Color by LogP, MW, or activity if known

Why it's good for your CV: Shows unsupervised learning, cheminformatics clustering, and visualization skills.

---

3. Descriptor Sensitivity Study for a Target Class

Goal: Test how sensitive model performance is to different descriptor sets (e.g., 2D, fingerprints, physicochemical).

- Use same data (e.g., kinase inhibitors from ChEMBL)
- Compare:
  - RDKit descriptors (`rdkit.Chem.Descriptors`)
  - Morgan fingerprints
  - MACCS keys
  - Hybrid sets
- Evaluate with ML models
- Report which descriptor types work best for the dataset

Why it's good for your CV: Demonstrates critical thinking and ability to evaluate model + feature selection.

---

4. Build a Web App: Molecular Descriptor Calculator

Goal: Let users input SMILES and get back descriptors, MW, LogP, and predicted class.

- Use: RDKit + Streamlit or Flask
- Compute:
  - Physicochemical (MW, TPSA, H-bond donors)
  - Fingerprints (bit vector or SVG structure)
  - Drug-likeness (Lipinski rule violation count)
- Optional: save history or plot similarities

Why it's good for your CV: Adds front-end skill, and shows you can turn chemistry code into tools for others.

---

5. Compare Descriptor Similarity vs. Bioactivity Similarity

Goal: Explore cases where structural similarity fails to predict similar bioactivity — and vice versa.

- Compute Tanimoto similarity between molecules
- Compare to bioactivity similarity (difference in IC_{50} or binary labels)
- Identify "activity cliffs" — similar structure, different activity
- Visualize outliers with molecule images

Why it's good for your CV: Shows domain insight and critical analysis of when descriptors *fail*, not just succeed.

Additional Enhancements

- Add **visualizations** (e.g., molecule grids, 2D projections)
- Export figures or data to **PDF reports** or use **interactive dashboards**
- Include **GitHub repo** with README, clear dependencies (`requirements.txt`)
- Document **why each descriptor is meaningful** (interpretability matters)

CV Bullet Point Examples

Bad:

• Worked with RDKit on descriptor stuff

Good:

• Built a machine learning pipeline using RDKit descriptors to predict kinase inhibitor potency (ChEMBL; $R^2 = 0.79$)
• Developed a clustering model (Tanimoto + t-SNE) to identify novel scaffold clusters from ZINC15 library (n = 10,000)
• Created a Streamlit-based web app for real-time SMILES descriptor calculation and Lipinski rule evaluation