Diffusion-Based Molecular Generation for Mutant p53 Drug Design

Jeremy T. Zay[1,2], and Kelly M. Thayer[1,2,3]

[1]Department of Mathematics & Computer Science, Wesleyan University, Middletown, CT 06459, USA
[2]College of Integrative Sciences, Wesleyan University, Middletown, CT 06459, USA
[3]Molecular Biophysics Program, Wesleyan University, Middletown, CT 06459, USA
jzay@wesleyan.edu

ABSTRACT

Designing small molecules that can both bind to a target protein and induce an allosteric restoration of its function remains a central challenge in drug discovery. Such molecules could reactivate dysfunctional proteins across many diseases; in this work, we focus on mutant p53, a tumor-suppressor protein whose loss of function is implicated in over half of human cancers.[5]

Traditional approaches have relied on reinforcement learning and docking-based reward functions to optimize autoregressive molecular generation[5], but these methods can be unstable and limited in scope[1]. Recent advances in diffusion-based generative models, which was originally popularized in image generation[6], have shown strong potential for generating drug-like molecules directly within protein binding pockets.

In this work, we explore how diffusion frameworks can be adapted and extended for therapeutic design, with a focus on generating ligands for mutant p53. We investigate strategies for incorporating additional objectives into diffusion-based sampling, allowing the generative process to be steered toward molecules with not only favorable binding properties but also other desired functional outcomes. Our study demonstrates the potential of diffusion-guided molecular generation as a flexible platform for integrating multiple design criteria in the pursuit of targeted cancer therapeutics.

INTRODUCTION

Designing drug-like molecules that fit specific protein pockets is difficult due to the huge chemical search space and complex 3D interactions. Recent diffusion models have shown promise for 3D molecular generation. TargetDiff[1] extends diffusion modeling to target-conditioned molecule generation, producing ligands that fit a protein pocket directly while respecting 3D geometry.

The model alternates between updating hidden features and coordinates using an SE(3)-equivariant Graph Neural Network. This guarantees rotational and translational invariance, which is crucial for realistic 3D chemistry.[1]

The protein is kept fixed while the ligand coordinates evolve under the learned denoising dynamics.[1]

RESULTS

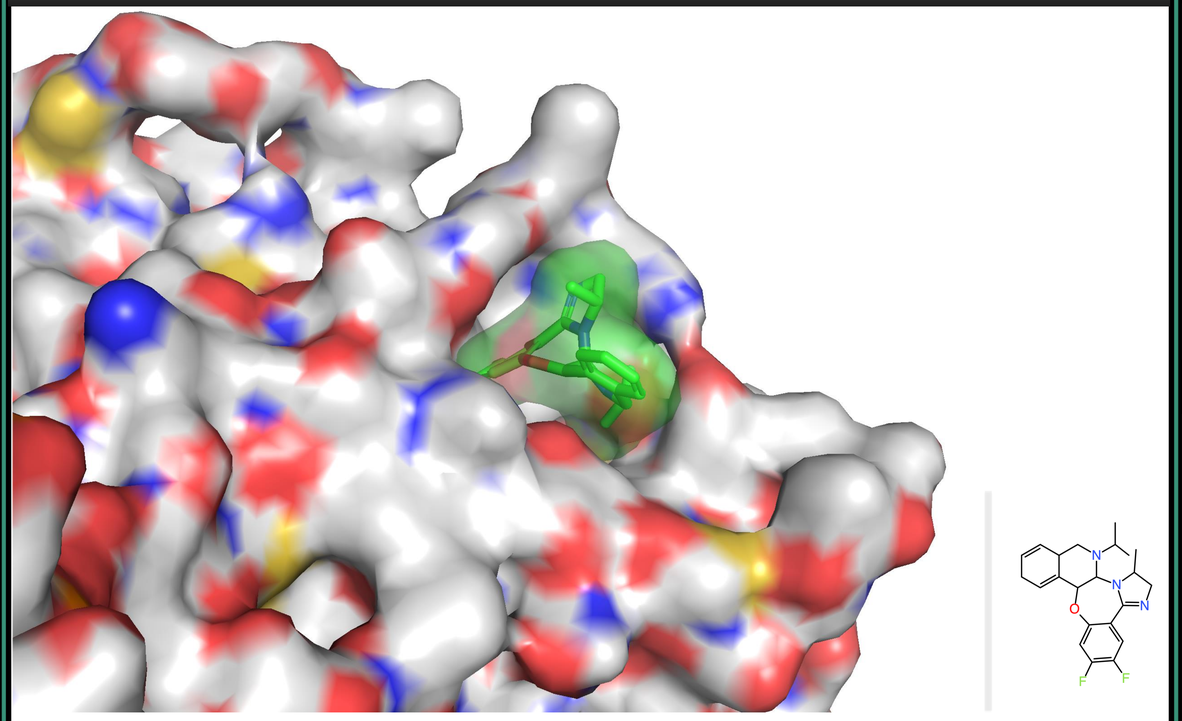
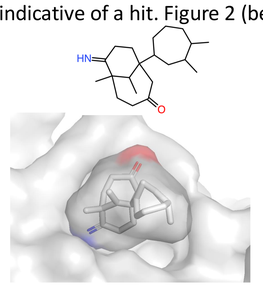
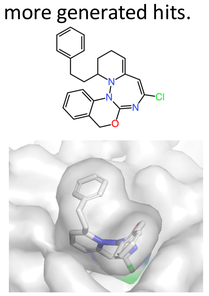


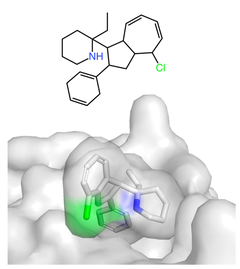
Figure 1, 2: Molecules Docked to Y220C Mutant p53. Figure 1 (above) is a p53[2] with a generated[1] molecule bound to it with a vina[3] score of -8.8 kcal/mol, a Quantitative Estimate of Druglikeness (QED) score of 0.69, and a Synthetic Accessibility (SA) score of 0.57, indicative of a hit. Figure 2 (below) contains more generated hits.



Vina: -8.9 kcal/mol
QED: 0.62, SA: 0.51



Vina: -8.5 kcal/mol
QED: 0.66, SA: 0.70



Vina: -8.4 kcal/mol
QED: 0.50, SA: 0.52

METHODS

Algorithm 1 Training Procedure of TargetDiff

Input: Protein-ligand binding dataset $\{\mathcal{P}, \mathcal{M}\}_{i=1}^{N}$, neural network ϕ_θ
while ϕ_θ not converge **do**
  Sample diffusion time $t \in \mathcal{U}(0, \ldots, T)$
  Move the complex to make CoM of protein atoms zero
  Perturb $\mathbf{x}_0$ to obtain $\mathbf{x}_t$: $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + (1 - \bar{\alpha}_t)\epsilon$, where $\epsilon \in \mathcal{N}(0, \boldsymbol{I})$
  Perturb $\mathbf{v}_0$ to obtain $\mathbf{v}_t$:
    $\log \boldsymbol{c} = \log(\bar{\alpha}_t \mathbf{v}_0 + (1 - \bar{\alpha}_t)/K)$
    $\mathbf{v}_t = \texttt{one\_hot}(\arg\max_i [g_i + \log c_i])$, where $g \sim \text{Gumbel}(0, 1)$
  Predict $[\hat{\mathbf{x}}_0, \hat{\mathbf{v}}_0]$ from $[\mathbf{x}_t, \mathbf{v}_t]$ with ϕ_θ: $[\hat{\mathbf{x}}_0, \hat{\mathbf{v}}_0] = \phi_\theta([\mathbf{x}_t, \mathbf{v}_t], t, \mathcal{P})$
  Compute the posterior atom types $\boldsymbol{c}(\mathbf{v}_t, \mathbf{v}_0)$ and $\boldsymbol{c}(\mathbf{v}_t, \hat{\mathbf{v}}_0)$ according to equation 4
  Compute the unweighted MSE loss on atom coordinates and the KL loss on posterior atom types: $L = \|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|^2 + \alpha\, \text{KL}(\boldsymbol{c}(\mathbf{v}_t, \mathbf{v}_0) \parallel \boldsymbol{c}(\mathbf{v}_t, \hat{\mathbf{v}}_0))$
  Update θ by minimizing L
end while

Algorithm 2 Sampling Procedure of TargetDiff

Input: The protein binding site $\mathcal{P}$, the learned model ϕ_θ.
Output: Generated ligand molecule $\mathcal{M}$ that binds to the protein pocket.
Sample the number of atoms in $\mathcal{M}$ based on a prior distribution conditioned on the pocket size
Move CoM of protein atoms to zero
Sample initial molecular atom coordinates $\mathbf{x}_T$ and atom types $\mathbf{v}_T$:
  $\mathbf{x}_T \in \mathcal{N}(0, \boldsymbol{I})$
  $\mathbf{v}_T = \texttt{one\_hot}(\arg\max_i g_i)$, where $g \sim \text{Gumbel}(0, 1)$
for t in $T, T-1, \ldots, 1$ **do**
  Predict $[\hat{\mathbf{x}}_0, \hat{\mathbf{v}}_0]$ from $[\mathbf{x}_t, \mathbf{v}_t]$ with ϕ_θ: $[\hat{\mathbf{x}}_0, \hat{\mathbf{v}}_0] = \phi_\theta([\mathbf{x}_t, \mathbf{v}_t], t, \mathcal{P})$
  Sample $\mathbf{x}_{t-1}$ from the posterior $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \hat{\mathbf{x}}_0)$ according to equation 4
  Sample $\mathbf{v}_{t-1}$ from the posterior $p_\theta(\mathbf{v}_{t-1}|\mathbf{v}_t, \hat{\mathbf{v}}_0)$ according to equation 4
end for

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}) \qquad q(\mathbf{v}_{t-1}|\mathbf{v}_t, \mathbf{v}_0) = \mathcal{C}(\mathbf{v}_{t-1}|\tilde{\boldsymbol{c}}_t(\mathbf{v}_t, \mathbf{v}_0)). \qquad (4)$$

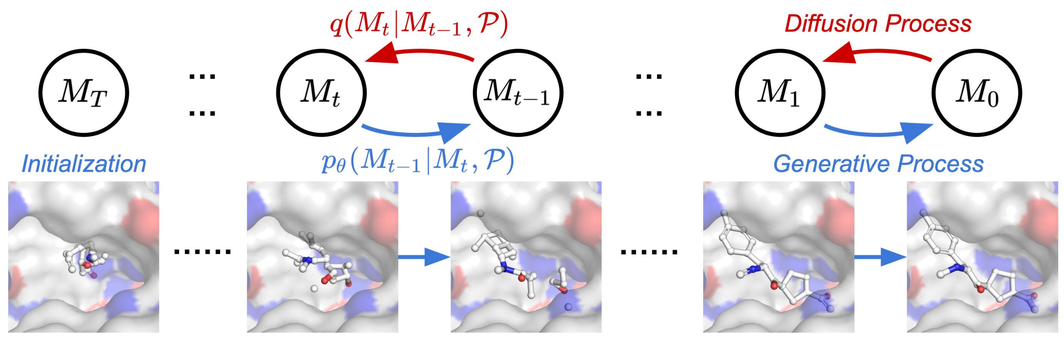


Figure 3: Schematic of the diffusion process. Noise is added to both coordinates and atom types during training, and denoised step-by-step during sampling using an equivariant GNN that jointly encodes protein–ligand interactions[1].

← **Algorithm 1, 2:** TargetDiff training and sampling procedures showing how noisy coordinates and atom types are denoised stepwise[1].

CONCLUSIONS

We generated 10,000 candidate molecules using our TargetDiff-based pipeline and evaluated their binding to Y220C p53 with AutoDock Vina
Most predicted binders scored between –7 and –5 kcal/mol, with 33 achieving ≤ –8.0 kcal/mol.
Using literature thresholds (–5 inclusive, –7 stringent), many of these exceed stringent binding criteria, indicating that the diffusion model reliably proposes high-affinity candidates.
Our top binder (–8.9 kcal/mol) nearly matches the strongest non-covalent Y220C ligand reported in BindingDB (–9.6 kcal/mol), despite those literature actives being hand-designed by medicinal-chemistry groups. In contrast, our fully automated pipeline produces comparably strong candidates efficiently (~15 min / 100 molecules), enabling rapid iteration and screening.

FUTURE DIRECTIONS

TargetDiff ⟶ p53-Restoration-Guided TargetDiff

One limitation of current SOTA 3D generative models is that they design binders, but none optimize for functional restoration of mutant p53. We will extend TargetDiff with differentiable classifier guidance to bias generation toward molecules that both bind to p53 and restore its wild-type activity.

At each reverse step, the model will adjust the mean of the coordinate Gaussian and the logits of atom types using the gradient of a restoration score $f_t(x_t, z_t, P)$:

$$\mu_{\text{guided}} = \mu_\theta(x_t, v_t) + s_x\, \beta_t\, \nabla_{x_t} f_t(x_t, z_t, P),$$
$$z_{\text{guided}} = z_t + s_v\, \nabla_{z_t} f_t(x_t, z_t, P).$$

We will train a graph neural network on MD simulations of Y220C p53 to predict the hydrogen-bond restoration Y220C with the docked drug compared to the wild-type p53, taking as input the noisy molecule at each diffusion timestep, the timestep t, and the embedded electrostatic network graph of p53 constructed using Node2vec.

This guidance will steer sampling toward conformations predicted to improve p53 binding affinity and functional restoration; however, we will need to show that the guidance function preserves SE(3)-equivariance and retains diffusion stability.

Ultimately, this approach will enable goal-directed generation of therapeutics that are physically plausible, pocket-specific, and functionally reparative.

ACKNOWLEDGEMENTS

We gratefully acknowledge Wesleyan's Molecules to Medicine Consortium, especially Prof. David L. Beveridge, Prof. Michael P Weir, and Dr. Matthew Young. We thank Guanlue Li, PhD at UHH, for fruitful discussion. We thank all members of the Thayer Lab for their inspiration and support. We would like to thank Mark Han for access to his trajectories, and Theo Sternlieb for earlier work related to the project. This work was supported by NIH R15 GM128102-01, CNS-0619508, CNS-095985, and NSF CHE-2320718.

REFERENCES

1. Guan et al., (2023).TargetDiff: Generating 3D Molecules for Protein Binding with SE(3)-Equivariant Diffusion (arXiv:2303.03543).
2. Han, S. M., & Thayer, K. M. (2023). Reconnaissance of allostery via the restoration of native p53 DNA-binding domain dynamics in Y220C mutant p53 tumor suppressor protein. *ACS Omega*. https://doi.org/10.1021/acsomega.3c08509.
3. Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry, 31*(2), 455-461.
4. Ozaki, T, & Nakagawara, A. (2011). Role of p53 in Cell Death and Human Cancers. https://pmc.ncbi.nlm.nih.gov/articles/PMC3756401/.
5. Peng, X., Luo, S., Zhao, X., Wu, J., & Ma, J. (2022). Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets. arXiv preprint arXiv:2205.07249.
6. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2006.11239.
7. Baud MGJ et al., 2018. Aminobenzothiazole derivatives stabilize the thermolabile p53 cancer mutant Y220C... Eur J Med Chem 152:101–114.
8. Wong et al., (2022). Benchmarking AlphaFold-enabled molecular docking predictions for antibiotic discovery. Molecular Systems Biology, 18(9), e11081. doi:10.15252/msb.202211081.