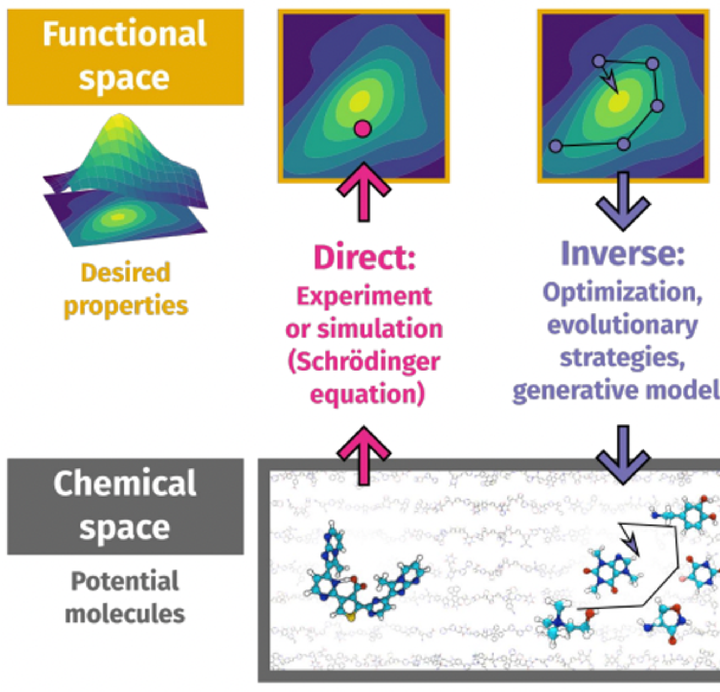
Pareto-guided Diffusion Model for Protein Design

Yinghua Yao, CFAR & IHPC, A*STAR, Singapore

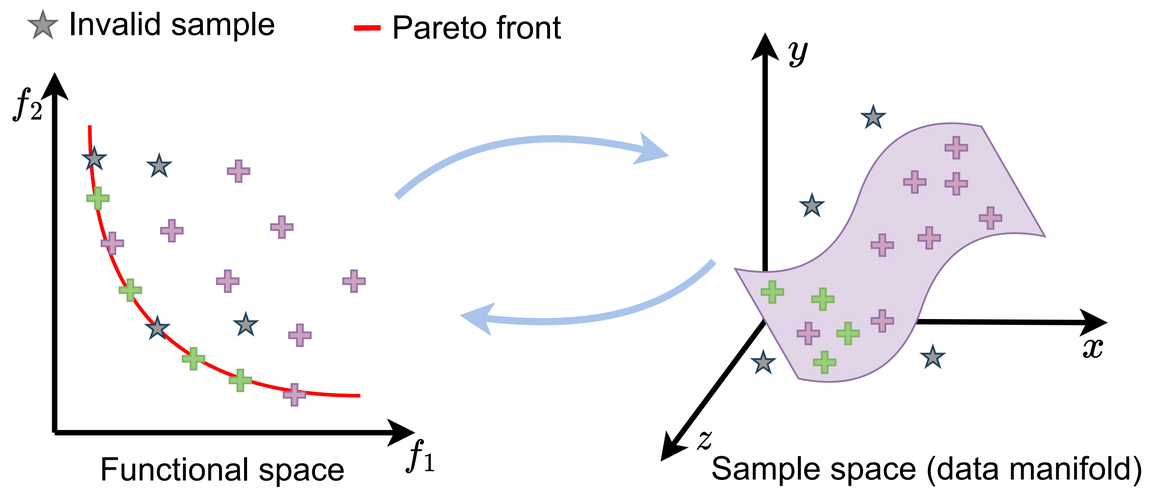
Background: Inverse (Generative) Design



Credits to Sanchez-Lengeling, B., & Aspuru-Guzik, A., 2018

- What we want can only be measured on the functional space.
- What we can manipulate is in the chemical space.
- Current main streams are *single-objective optimization/generation*.

Multi-Objective Generation (MOG) for Protein Design



- Protein data: low-dimensional manifold of high-dimensional space
- Pareto front: trade-off between multiple property objectives

Superiority of MOG over existing *Multi-Objective Optimization* (MOO)

	objectives	decision/data space	generation quality
MOO	$F(x) = [f_1(x), f_2(x), \ldots, f_m(x)]$	$x \in \mathbb{R}^d$	✗
MOG		$x \in \mathcal{X}, \mathcal{X} \subset \mathbb{R}^d$	✓

Constrained Optimization for MOG

Let p_0 be the distribution of samples on Pareto front, and $p_\theta(x)$ be the target data distribution, our constrained optimization for MOG can be formulated as

$$\min_\theta D\left[q_{\text{data}}(x)||p_\theta(x)\right] \quad s.t.\ D\left[p_0(x)||p_\theta(x)\right] \leq \varepsilon.$$

- Data quality: minimize the KL divergence between the training data distribution q_{data} and the generated data distribution p_θ.
- Pareto optimality: p_θ constrained to be close to the distribution of Pareto solutions p_0.

Optimize $D\left[q_{\text{data}}(x)||p_\theta(x)\right]$ with diffusion model:

$$x_{t-1} = x_t - \eta_t \epsilon^*_\theta(x_t, t) + \sqrt{2\eta_t} z.$$

Optimize $D\left[p_0(x)||p_\theta(x)\right]$ with multiple gradient descent:

$$x_{t-1} = x_t - \eta \nabla F(x_t) + \sqrt{2\eta} z,$$

Pareto-guided Diffusion Model

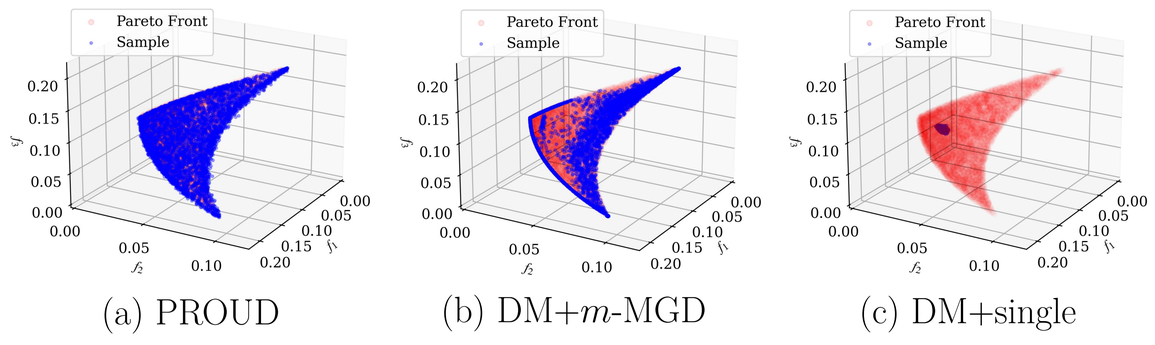
Overall reverse diffusion: $x_{t-1} = x_t - \eta_t g(x_t) + \sqrt{2\eta_t} z$, where

$$g(x_t) = \arg\min_g \frac{1}{2}\|g - \epsilon^*_\theta(x_t, t)\|^2 \quad s.t.\ \nabla f_i(x)^T g \geq \phi_t, \quad \forall i = 1, 2, \ldots, m,$$

$$\phi_t = \begin{cases} \alpha \|\nabla F(x_t)\| & \text{if } \|\nabla F(x_t)\| > e \\ -\infty & \text{otherwise} \end{cases},$$

1. Constraint violation: optimize $g(x_t)$ to decrease all objectives
2. Constraint violation: optimize $g(x_t)$ to keep data quality as much as possible
3. Constraint satisfaction: optimize $g(x_t)$ to keep data quality

Effective in Covering the Pareto Front



(a) PROUD (b) DM+m-MGD (c) DM+single

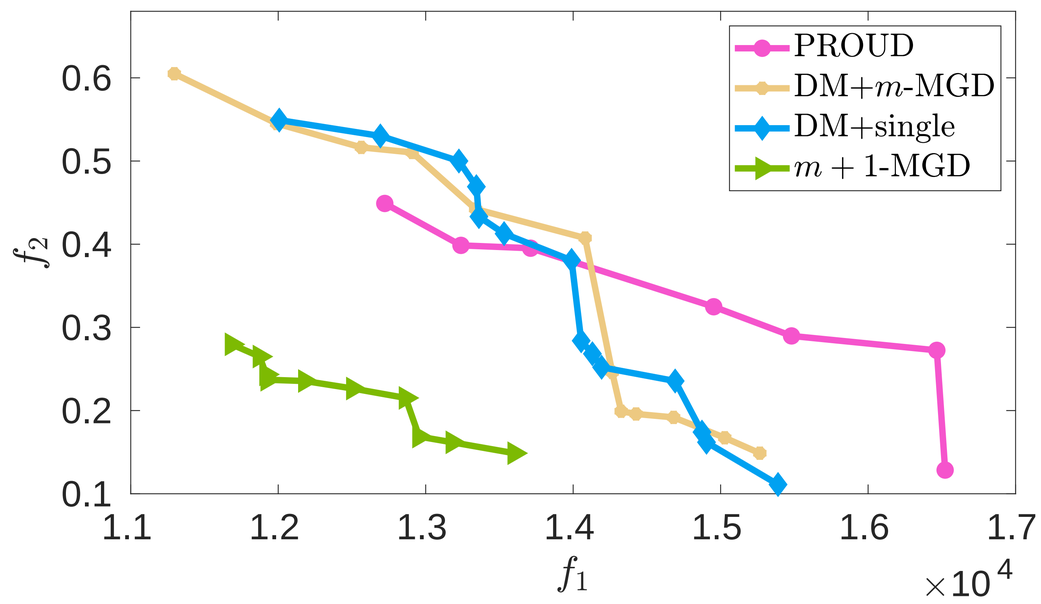
Experiments on Protein Sequences

Dataset: paired Observed Antibody Space (pOAS) dataset comprised of 90, 990 antibody sequences.

Metrics: Hypervolume (HV) for Pareto front approximation, the log-likelihood assigned by ProtGPT for the quality of generated protein sequences.

Multiple desired properties:

- $f_1(x)$: solvent accessible surface area (SASA) of the protein structure
- $f_2(x)$: percentage of beta sheets (%Sheets)



Method	HV↑ (Pareto optimality)	ProtGPT↑ (data quality)
PROUD (ours)	**2472.55±60.15**	**-645.93±0.99**
DM+m-MGD	2289.61±65.12	-692.80±0.34
DM+single	2302.21±58.25	-682.26±0.49
$m+1$-MGD	838.74±14.08	-662.86±0.76

Paper

Centre for Frontier AI Research

CFAR