Table 1: **[A1]** Additional baseline results for APHYNITY and Genetic programming for symbolic regression. APHYNITY is implemented using domain-specific expert models as defined in App H.8 combined with a 3-layer MLP, with the same hyper-parameters from [R2]. Genetic programming is implemented using the implementation and hyper-parameters from [R6]. Reporting the test prediction MSE ($\mathcal{T}_{MSE}$) of the produced system models on held-out test datasets across all benchmark datasets. HDTwinGen achieves the lowest test prediction error. The results are averaged over ten random seeds, with $\pm$ indicating 95% confidence intervals.

| Method	Lung Cancer $\mathcal{T}_{MSE}\downarrow$	Lung Cancer (with Chemo.) $\mathcal{T}_{MSE}\downarrow$	Lung Cancer (with Chemo. & Radio.) $\mathcal{T}_{MSE}\downarrow$	Hare-Lynx $\mathcal{T}_{MSE}\downarrow$	Plankton Microcosm $\mathcal{T}_{MSE}\downarrow$	COVID-19 $\mathcal{T}_{MSE}\downarrow$
DyNODE	327±5.8	52±47.1	16.3±5.58	439±0	0.00036±0.00078	74±2.36
SINDy	327±5.79	11.8±0.395	13.7±0.573	388±4.29e-14	0.00135±0	93.4±0.458
RNN	1.17e+06±3.08e+04	708±86.1	136±5.6	3.71e+03±3.39e+03	0.0281±0.0406	1.38e+04±1.65e+03
Transformer	7.48±1.06	0.348±0.0618	0.216±0.0345	716±42.5	3.69e-05±1.83e-05	0.309±0.222
APHYNITY	9.06±1.37	81.6±81.3	1.21e+03±1.69e+03	321±12.6	4.21e-05±3.45e-05	88.8±9.97
Genetic programming	184±47	187±30.4	152±30.6	504±34.1	0.00928±0.00539	7.86±1.89
ZeroShot	5.45e+03±6.71e+03	292±80.2	5.81e+03±4.02e+03	338±0	0.325±0.242	2.31e+03±2.24e+03
ZeroOptim	216±172	31.2±45	6.08±7.9	353±0	0.0132±0.00116	7.88±0.0414
HDTwinGen	**4.41±8.07**	**0.0889±0.0453**	**0.131±0.198**	**291±30.3**	**2.51e-06±2.2e-06**	**1.72±2.28**
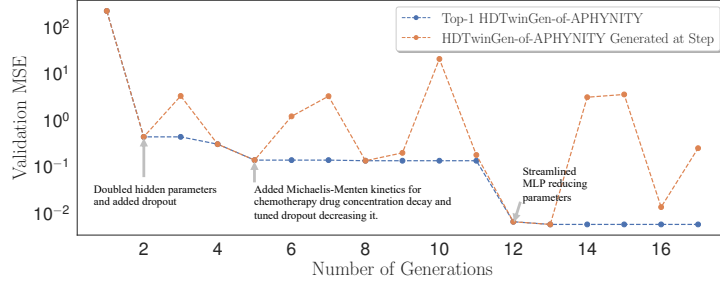
Figure 1: **[A2]** HDTwinGen can effectively evolve an existing hybrid APHYNITY model. Validation MSE of the hybrid model generated in each iteration, showing the Pareto-front of the best generated hybrid model (Top-1 HDTwinGen-of-APHYNITY) and the generated HDTwinGen-of-APHYNITY per generation step—additionally with a few of the hybrid models labeled with their model descriptions. HDTwinGen can efficiently understand, modify, and hence *evolve* the hybrid model to achieve a better-fitting model; this experiment is on the Lung Cancer (with Chemo. & Radio.) dataset.
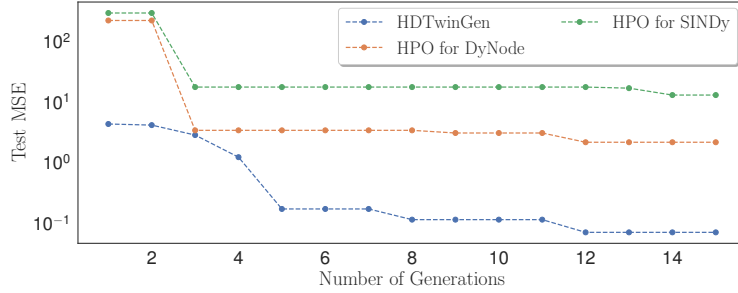

Figure 2: **[A3]** Bayesian Optimization of baseline hyperparameters over the same number of model generations to provide baselines with a comparable compute budget and automated model improvement comparison. Test MSE of the best-performing model against each generation step, up to 15 iterations. HDTwinGen generates better-performing hybrid models for a fixed model budget. The experiment is performed on the Lung Cancer (with Chemo. & Radio.) dataset. Hyper-parameter search space for the baselines is detailed in Table 3, and the Bayesian optimization algorithm used for HPT is a Tree-structured Parzen Estimator.

Table 2: SINDy grid search over polynomial library degree.

Method	Lung Cancer (with Chemo. & Radio.) $\mathcal{T}_{MSE}\downarrow$
SINDy (polynomial_library_degree=1)	132±7.02
SINDy (polynomial_library_degree=2)	13.9±1.12
SINDy (polynomial_library_degree=3)	14±1.11
SINDy (polynomial_library_degree=4)	14±1.13

Table 3: BO HPT Search Space for Figure 2.

Method	Hyperparameter Search Space	Best Params
DyNode learning_rate	[1e-5, 1e-1]	0.0123
DyNode weight_decay	[1e-5, 1e-1]	0.00029
DyNode hidden_dim	[32, 1024]	788
DyNode model_activation	[tanh, silu, ELU]	tanh
DyNode model_initialization	[xavier, normal]	normal
SINDy polynomial_library_degree	[1, 3]	2
SINDy polynomial_library_interaction_only	[True, False]	True
SINDy threshold	[1e-5, 1e-1]	0.0194
SINDy alpha	[1e-5, 1e-1]	0.0015