Supplemental Material

A Deep Learning Approach to Searching Property Spaces of Materials

Robert J. Appleton[1], Brian C. Barnes[2], Steven F. Son[3], Alejandro Strachan[1]

[1] School of Materials Engineering and Birck Nanotechnology Center, Purdue University, West Lafayette, Indiana 47907, USA

[2] U.S. Army Combat Capabilities Development Command Army Research Laboratory, Aberdeen Proving Ground, Maryland 21005, USA

[3] School of Mechanical Engineering, Purdue University, West Lafayette, Indiana 47907, USA

S1. Model Hyperparameters and Training Details

Table S1. *Summary of hyperparameters to train D-MPNN. All unlisted hyperparameters are the defaults. Models were developed with chemprop 1.7.0.*[1,2]

Parameter	Value(s)
Max epochs	100
Batch size	16
Train-validation splits	0.9, 0.1
Number of message-passing steps	5
Dimensionality of MPNN hidden layers	1200
Dropout probability	0.25
Number of layers in FFN	2
Dimensionality of FFN hidden layers	1200
Targets (scaled with standard scalar)	T_m, T_b, T_d, $\log_{10} P_v$
Target weights	1, 5, 35, 10

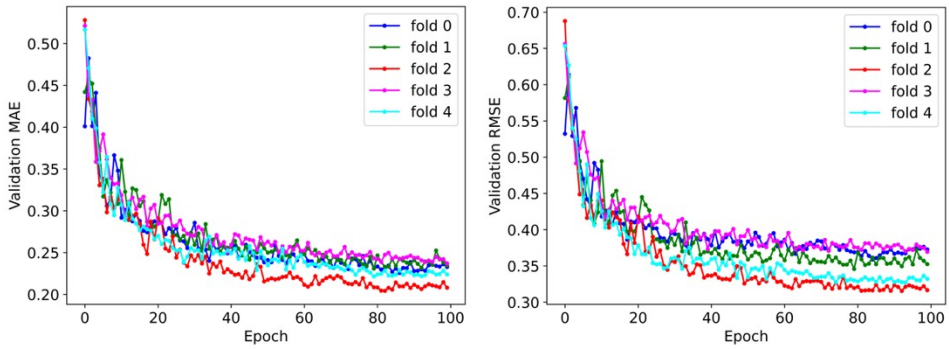

Figure S1. *Learning curves for each model: validation mean absolute error (**a**) and validation root mean squared error (**b**).*

S2. Trivial Solutions in Optimization Problems

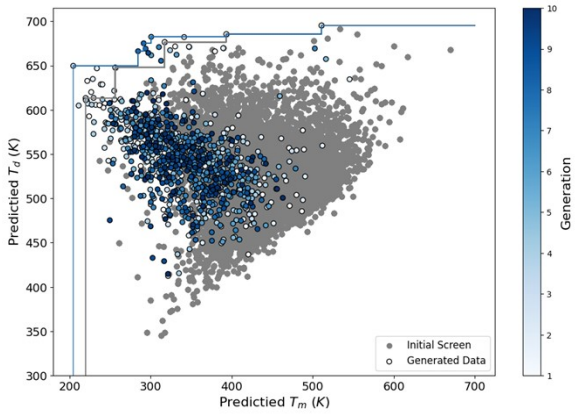

Figure S2. *Distribution of predicted decomposition temperature and predicted melting temperature. This preliminary result highlights that Pareto optimization alone can lead to trivial solutions. In this example, we find that the GA finds it easier to generate thermally stable liquids and gases, as shown by the number of points with melting temperatures below room temperature.*

S3. Parameters for Genetic Algorithm

Table S2. *Summary of parameters used in the JANUS genetic algorithm for all experiments. Full details of each parameter can be referenced here.[3] All unlisted parameters are the defaults.*

Parameter	Value(s)
Crossover number of random samples	5 (sampling experiments) 1*, 5 (hyperparameter experiments)
Exploit number of mutations	5 (sampling experiments) 10% of generation size, 400* (hyperparameter experiments)
Exploit number of random samples	5 (sampling experiments) 10% of generation size, 400* (hyperparameter experiments)
Generation size	50 (sampling experiments) 50, 100, 200 (hyperparameter experiments)
Generations	10 (sampling experiments) 25 (hyperparameter experiments)
Top molecules	5

* Indicates a default parameter for JANUS.

S4. Sampling Experiments

S4.1. Liquid Stability Range and Distance from Ideal Melting Temperature

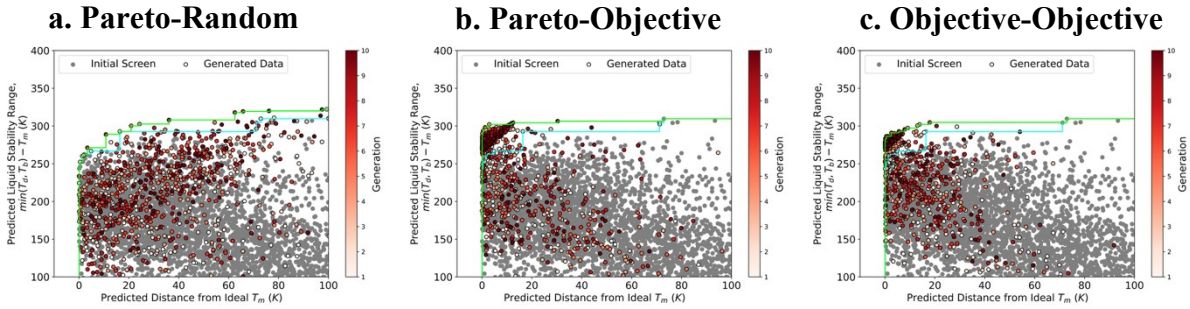

Figure S3. *Distribution of predicted liquid stability ranges and distances from ideal melting temperature for CHNO molecules from literature (grey) and generated (red). The Pareto front for the initial population is represented by the cyan step function and the Pareto front for the final population is represented by the lime green step function. The results using random down-sampling in the explorative population and distance from the Pareto front as the fitness function are shown in (**a**). The results using the objective function for down-sampling in the explorative population and distance from the Pareto front as the fitness function are shown in (**b**). The results using the objective function for down-sampling and the fitness function are shown in (**c**). Only a single experiment is shown in each plot, however, each experiment was repeated 5 times for statistics.*

S4.2. Liquid Stability Range and Vapor Pressure

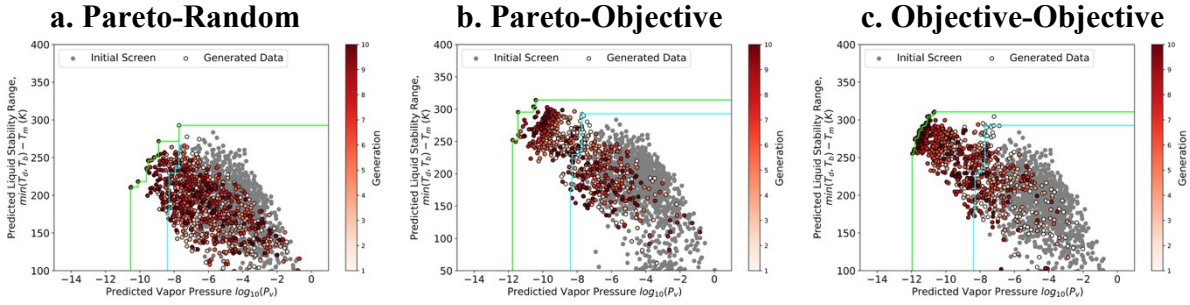

Figure S4. *Distribution of predicted liquid stability ranges and vapor pressures for CHNO molecules from literature (grey) and generated (red). The Pareto front for the initial population is represented by the cyan step function and the Pareto front for the final population is represented by the lime green step function. The results using random down-sampling in the explorative population and distance from the Pareto front as the fitness function are shown in (**a**). The results using the objective function for down-sampling in the explorative population and distance from the Pareto front as the fitness function are shown in (**b**). The results using the objective function for down-sampling and the fitness function are shown in (**c**). Only a single experiment is shown in each plot, however, each experiment was repeated 5 times for statistics.*

S5. Hyperparameter Experiments

S5.1. Liquid Stability Range and Distance from Ideal Melting Temperature

Table S3. *Results of the hyperparameter experiments for optimizing* ΔT_{liq} *and* ΔT_{im}*. Each experiment was evolved for 25 generations. In the table below,* N_{pop} *is the generation size,* N_{cross} *is the number of crossovers in the exploration population,* N_{mut} *is the number of mutations and random samples for the exploitation population,* N_{mols} *eval is the number of unique molecules evaluated,* N_{mols} *pop is the number of unique molecules that were selected into the population, and* N_{mols} *melt-cast is the number of unique molecules from the population that met the criteria to be considered melt-castable.*

N_{pop}	N_{cross}	N_{mut}	N_{mols} eval	N_{mols} pop	N_{mols} melt-cast	Wall time
50	5	5	305944	2443	1810	3D-19HR
50	1	400	618518	1475	1322	7D-16HR
100	5	10	523796	4482	3460	5D-16HR
100	1	400	718055	4244	3954	9D-16HR
200	5	20	705991	8879	7356	8D-07HR
200	1	400	842972	8263	6790	9D-06HR

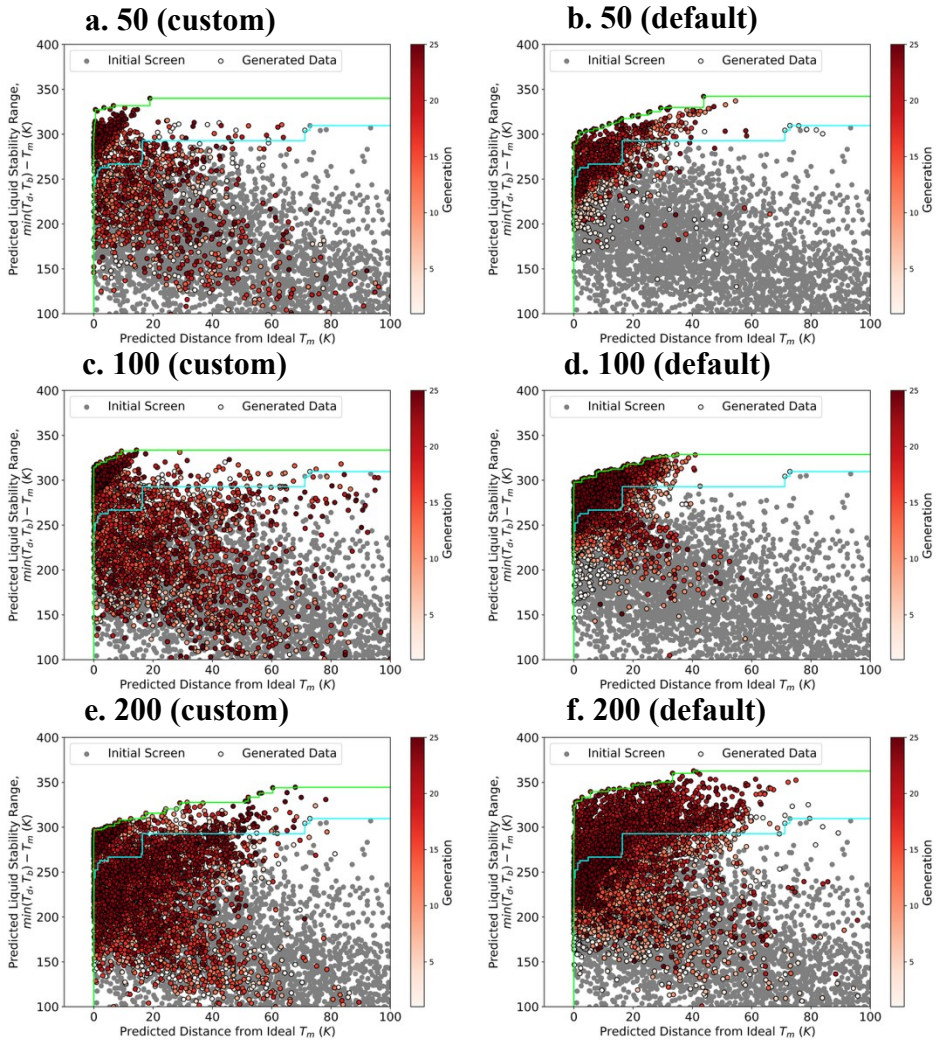

Figure S5. *Distribution of predicted liquid stability ranges and distances from ideal melting temperature for CHNO molecules from literature (grey) and generated (red). The Pareto front for the initial population is represented by the cyan step function and the Pareto front for the final population is represented by the lime green step function. The results are shown for a population size of 50 (**a,b**), 100 (**c,d**), and 200 (**e,f**).*

S5.2. Liquid Stability Range and Vapor Pressure

Table S4. *Same as Table S3, except these are the results of hyperparameter experiments for optimizing* ΔT_{liq} *and* $\log_{10}(P_v)$*. Each experiment was evolved for 25 generations.*

N_{pop}	N_{cross}	N_{mut}	N_{mols} eval	N_{mols} pop	N_{mols} melt-cast	Wall time
50	5	5	214015	2205	2179	2D-02HR
50	1	400	605562	1770	1770	7D-14HR
100	5	10	328037	4435	4323	4D-10HR
100	1	400	619238	3862	3862	8D-10HR
200	5	20	678909	8980	8748	6D-02HR
200	1	400	943894	8308	8305	12D-12HR

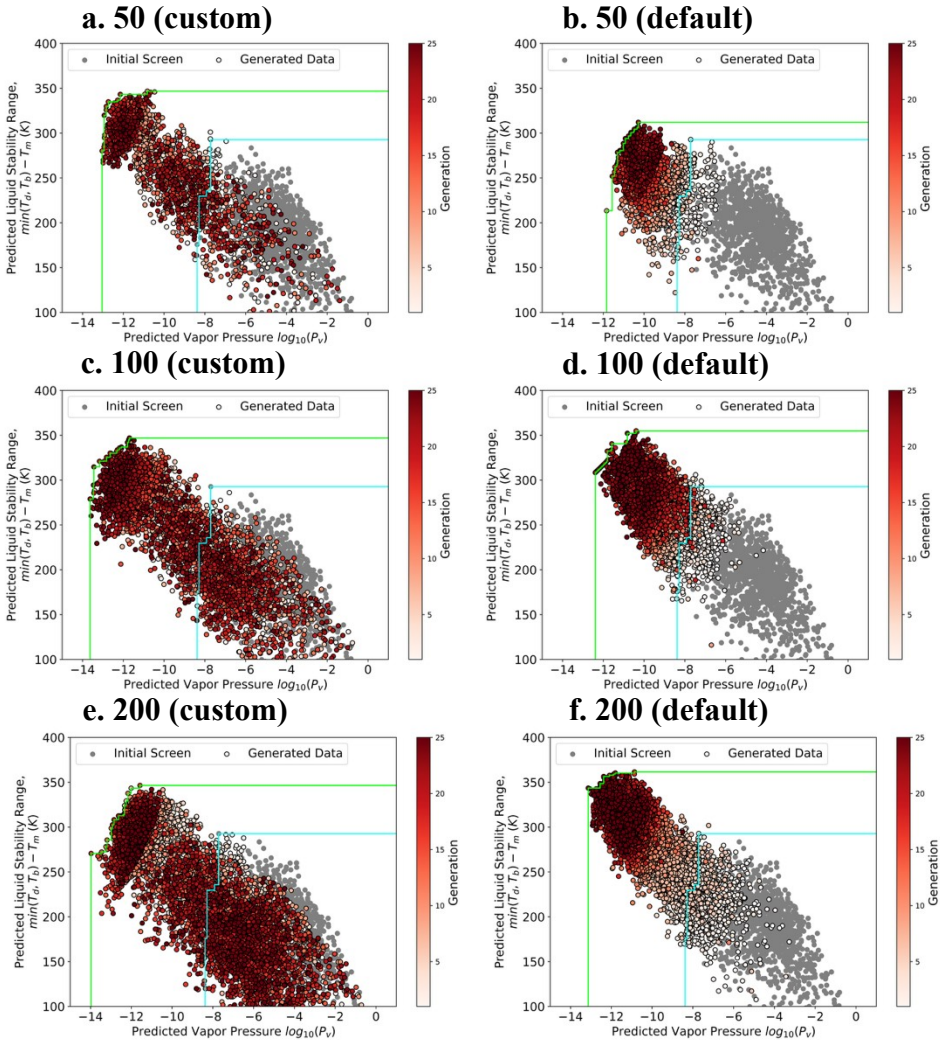

Figure S6. *Distribution of predicted liquid stability ranges and vapor pressure for CHNO molecules from literature (grey) and generated (red). The Pareto front for the initial population is represented by the cyan step function and the Pareto front for the final population is represented by the lime green step function. The results are shown for a population size of 50 (**a,b**), 100 (**c,d**), and 200 (**e,f**).*

S6. Generated Molecules

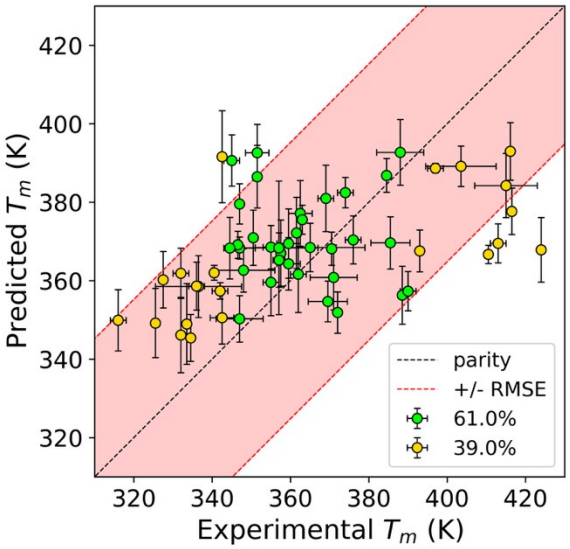

Figure S7. *Available experimental melting temperatures for the rediscovered molecules compared to the ML predictions. The vertical error bars correspond to the standard deviation of the predictions across the 5 models. The horizontal error bars correspond to the standard deviation of the experimental measurements. The red region captures +/- 35.2K (the evaluated RMSE of the model). The green points have a melting temperature within the range to be melted via steam heating, and the yellow points are outside the range.*

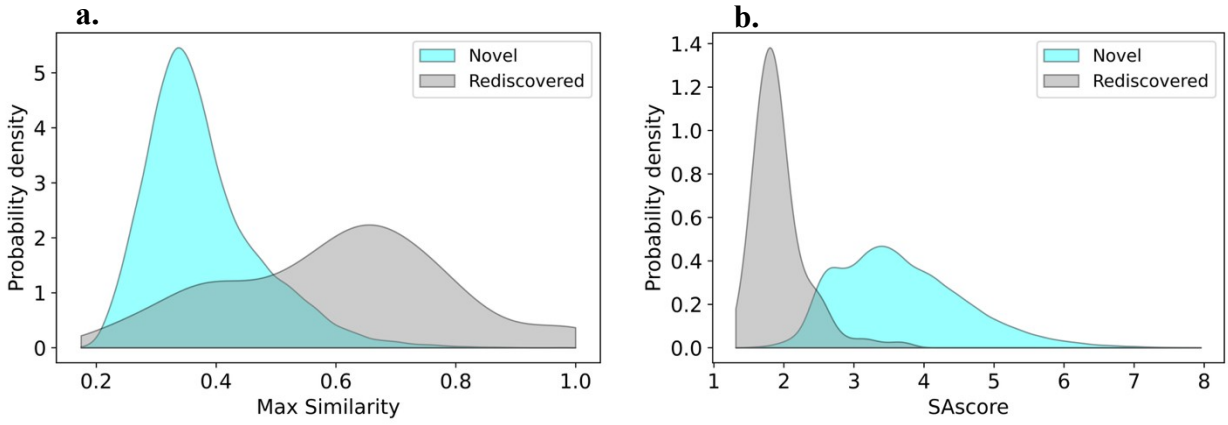

Figure S8. *Distributions of maximum similarity **(a)** and SAscore **(b)** for the novel generated melt-cast candidates (cyan) and the rediscovered molecules (grey).*

Supplemental References

1. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K. & Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J Chem Inf Model* 59, 3370–3388 (2019).

2. Heid, E., Greenman, K. P., Chung, Y., Li, S. C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H. & McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J Chem Inf Model* 64, 9–17 (2024).

3. Nigam, A. K., Pollice, R. & Aspuru-Guzik, A. Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. *Digital Discovery* 1, 390–404 (2022).