Appendix for "Discovering Infinite Recursive Conjectures through Genetic Programming"

This document provides the supplementary information to enhance the clarity and completeness of our main document. The supplementary materials are structured into three key sections. In Section A, we present a comprehensive table of acronyms and their definitions used throughout the main document, addressing potential terminology barriers for readers from diverse scientific backgrounds. Section B contains additional experimental results and corresponding analyses that could not be accommodated in the main text due to space limitations. These supplementary materials offer deeper insights into our methodology and findings, further supporting the conclusions drawn in the primary manuscript.

APPENDIX A
LIST OF ACRONYMS

In this section, we provide a comprehensive compilation of all acronyms and technical abbreviations used throughout the manuscript. This reference table is designed to enhance readability and accessibility for readers across different disciplines, particularly those who may be new to the field. The acronyms are presented in alphabetical order with their complete expansions and, where appropriate, brief descriptions of their significance in the context of our work.

TABLE A1
LIST OF ACRONYMS AND THEIR DEFINITIONS USED IN THE MAIN DOCUMENT.

Acronyms	Definitions
ADF	Automatically defined function
CL	Confidence level
CR	Convergence rate
DCE	Dual-chromosome encoding
DE	Differential evolution
GEP	Genetic expression programming
GP	Genetic programming
GR	Growth rate
ICE	Infinite conjecture explorer
LHS	Left-hand side
RHS	Right-hand side
SR	Success rate
TMO	Two-sided matching operator

APPENDIX B
ADDITIONAL EXPERIMENTAL RESULTS

This section presents the comprehensive experimental results for success rate (SR), confidence level (CL), and convergence rate (CR) metrics, complementing the growth rate (GR) analysis provided in the main text. While GR serves as our primary indicator of algorithmic performance due to its direct measurement of conjecture discovery efficiency, these three additional metrics offer valuable insights into other important aspects of algorithm behavior. SR quantifies the stability and reliability of the algorithm across multiple independent runs, CL measures the precision of the discovered conjectures, and CR characterizes the convergence behavior throughout the evolutionary process. Together, these metrics provide a more complete picture of how each algorithmic component contributes to overall performance beyond the production efficiency captured by GR alone.

TABLE B2
COMPARISON RESULTS OF THE SR. THE BEST VALUE AMONG THESE ALGORITHMS ON EACH TEST PROBLEM IS HIGHLIGHTED IN GRAY.

	GP	ADF-GEP	ICE
GIRCP1	52.00%	68.00%	100.00%
GIRCP2	46.00%	70.00%	100.00%
GIRCP3	54.00%	86.00%	100.00%
GIRCP4	64.00%	90.00%	100.00%
GIRCP5	52.00%	76.00%	98.00%
GIRCP6	52.00%	78.00%	100.00%

A. Results of the Comparative Experiment

This subsection analyzes the performance of ICE compared to GP and ADF-GEP based on the three complementary metrics of success SR, CL and CR. Table B2, Table B3 and Fig. B1 present the experimental results of GP, ADF-GEP, and ICE on six GIRCPs. These results are based on 50 independent runs and provide additional evaluation dimensions beyond the GR metric presented in the main text.

As shown in Table B2, ICE demonstrates exceptional stability and reliability, achieving a perfect 100% success rate on five of the six benchmark problems and 98% on GIRCP5. This indicates that ICE consistently produces meaningful conjectures across nearly all independent runs. In contrast, ADF-GEP achieves moderate success rates ranging from 68% to 90%, while GP exhibits the lowest success rates between 46% and 64%. The substantial performance gap between ICE and the other algorithms highlights the robustness of our proposed method in reliably generating valid conjectures, regardless of the specific problem characteristics or initial conditions. This consistent reliability represents a significant advancement over existing approaches, particularly for practitioners who require dependable conjecture generation capabilities.

Table B3 presents the mean and standard deviation of confidence level values, which measure the precision of the discovered conjectures. Statistical significance testing reveals that ICE outperforms GP on four problems (GIRCP1, GIRCP2, GIRCP3, and GIRCP5) and performs comparably on two (GIRCP4 and GIRCP6). When compared to ADF-

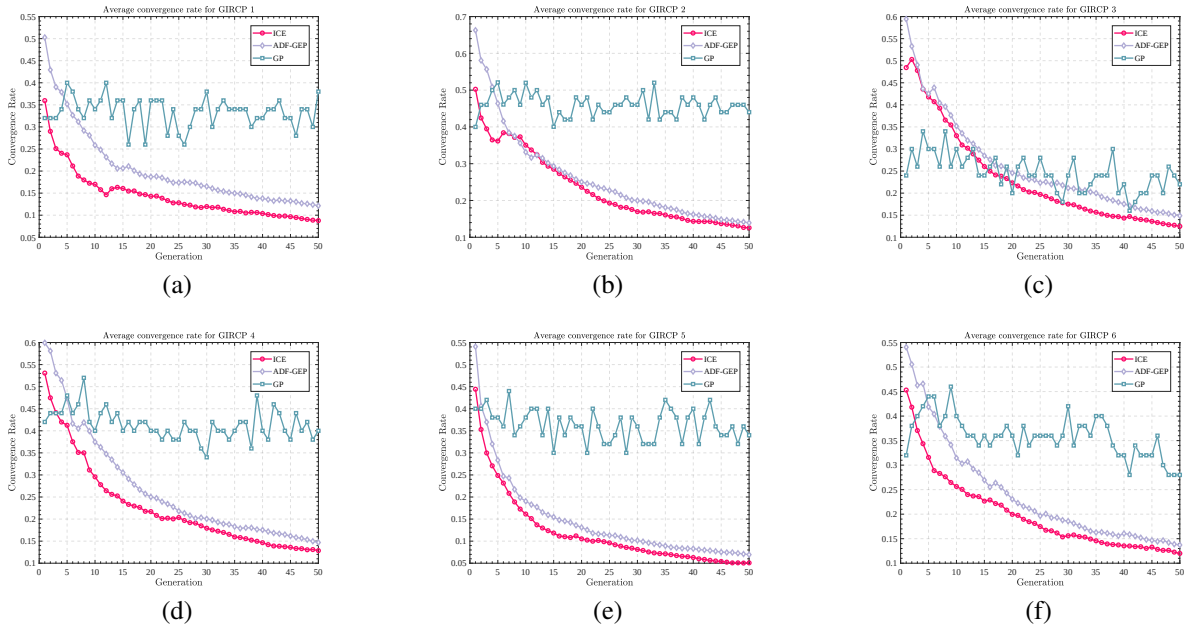

Fig. B1. Average CR of the three algorithms over 50 independent runs on six GIRCP problems. (a) GIRCP1, (b) GIRCP2, (c) GIRCP3, (d) GIRCP4, (e) GIRCP5, (f) GIRCP6.

TABLE B3
COMPARISON RESULTS OF THE MEAN VALUES AND STD OF CL.

	GP		ADF-GEP		ICE	
	Mean	Std	Mean	Std	Mean	Std
GIRCP1	2.82 $-$	2.51	4.03 $\approx$	2.80	4.92	2.30
GIRCP2	4.00 $-$	3.45	3.53 $-$	2.07	5.75	2.22
GIRCP3	3.42 $-$	2.54	4.77 $\approx$	2.19	4.86	1.76
GIRCP4	4.53 $\approx$	3.66	4.87 $\approx$	2.14	4.88	1.68
GIRCP5	4.73 $-$	4.08	3.91 $-$	1.20	5.99	1.45
GIRCP6	4.48 $\approx$	3.75	3.87 $-$	1.88	4.81	1.38
$+/\approx/-$	0/2/4		0/3/3			

Symbols $-, \approx$ and $+$ represent that the competitor is worse than, similar to, and better than ICE according to the t-test at $\alpha = 0.05$.

TABLE B4
COMPARISON RESULTS OF THE SR. THE BEST VALUE AMONG THESE ALGORITHMS ON EACH TEST PROBLEM IS HIGHLIGHTED IN GRAY.

	$N_O = 1$	Without ADF	Single-point crossover	ICE
GIRCP1	100.00%	64.00%	20.00%	100.00%
GIRCP2	98.00%	56.00%	16.00%	100.00%
GIRCP3	100.00%	76.00%	12.00%	100.00%
GIRCP4	100.00%	82.00%	20.00%	100.00%
GIRCP5	96.00%	56.00%	16.00%	98.00%
GIRCP6	100.00%	50.00%	20.00%	100.00%

GEP, ICE demonstrates significantly higher CL values on three problems (GIRCP2, GIRCP5, and GIRCP6) and comparable performance on the remaining three. The mean CL values for ICE range from 4.81 to 5.99, consistently higher than those of GP (2.82 to 4.73) and ADF-GEP (3.53 to 4.87). Moreover, ICE exhibits lower standard deviations across all benchmarks, indicating more consistent precision in the generated conjectures. These results demonstrate that ICE not only discovers conjectures more reliably but also produces conjectures with higher mathematical precision and consistency.

Fig. B1 illustrates the convergence behavior of the three algorithms across all six benchmark problems. ICE (red line) consistently achieves and maintains lower CR values throughout the evolutionary process compared to ADF-GEP (purple) and GP (teal). The convergence advantage of ICE becomes increasingly pronounced in later generations, particularly after generation 20. GP shows highly unstable convergence patterns with substantial fluctuations, reflected in the jagged blue curves across all problems. While ADF-GEP demonstrates more stable convergence than GP, it consistently lags behind ICE, especially in the later stages of evolution. This superior convergence behavior indicates that ICE more efficiently navigates the search space of potential conjectures, identifying higher-quality solutions earlier in the evolutionary process. The consistent downward trend of ICE's convergence curves across all benchmarks further demonstrates the effectiveness of its search mechanisms in progressively refining conjectures throughout the evolutionary process.

Together, these three metrics provide comprehensive evidence of ICE's superior performance beyond the GR results presented in the main text. The consistent excellence across multiple performance dimensions confirms that ICE represents a significant advancement in automated mathematical conjecture generation.

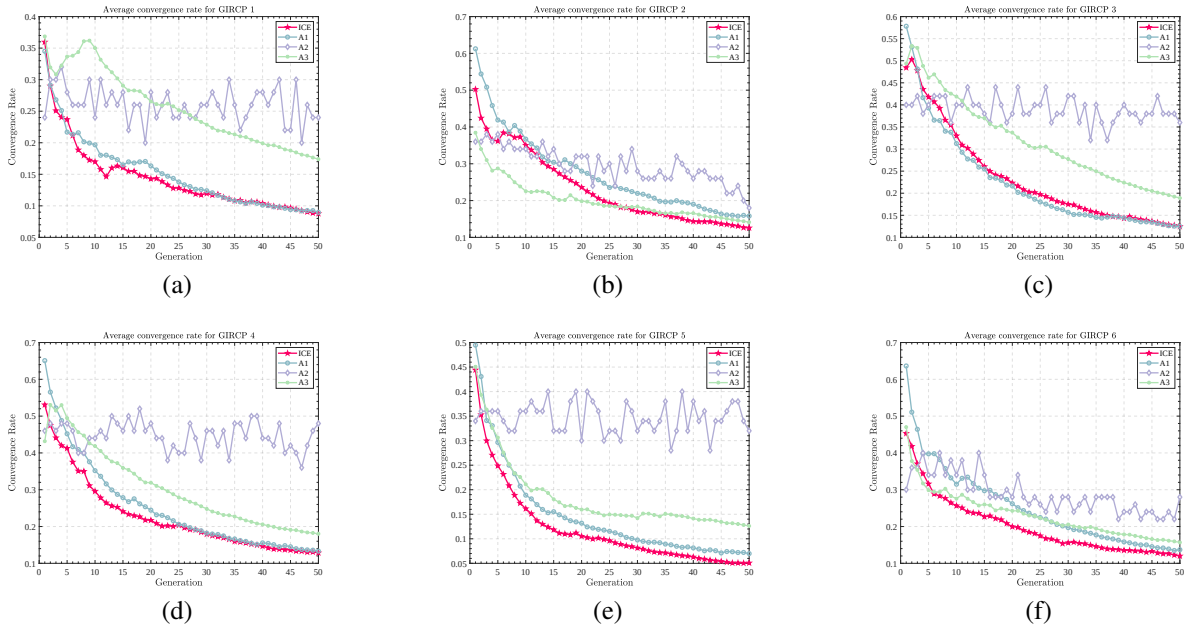

Fig. B2. Average CR of the four algorithms over 50 independent runs on six GIRCP problems under ablation settings. (a) GIRCP1, (b) GIRCP2, (c) GIRCP3, (d) GIRCP4, (e) GIRCP5, (f) GIRCP6. ICE (red) represents the baseline method. A1 (purple, removing pre-selection, $N_O = 1$) shows convergence patterns similar to ICE in later generations for some benchmarks but with higher variability. A2 (green, excluding ADF encoding in LHS population) generally shows improved convergence over A3 but underperforms ICE in most benchmarks. A3 (blue, replacing frequency-based DE with single-point crossover) demonstrates the highest CR values and greatest variability, indicating poorest convergence. These varied patterns highlight how different algorithmic components contribute to convergence behavior across different problem types.

TABLE B5
COMPARISON RESULTS OF THE MEAN VALUES AND STD OF CL IN ABLATION EXPERIMENTS.

Problem	A1		A2		A3		ICE	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
GIRCP1	4.60 $\approx$	1.08	2.46 $-$	1.37	1.14 $-$	0.35	4.92	2.30
GIRCP2	5.04 $\approx$	1.30	3.43 $-$	1.84	1.21 $-$	0.42	5.75	2.22
GIRCP3	5.27 $\approx$	1.35	2.90 $-$	1.58	1.08 $-$	0.27	4.86	1.76
GIRCP4	4.81 $\approx$	0.98	2.90 $-$	1.59	1.15 $-$	0.48	4.88	1.68
GIRCP5	4.15 $-$	1.07	2.94 $-$	1.59	1.18 $-$	0.48	5.99	1.45
GIRCP6	4.55 $\approx$	0.86	3.22 $-$	1.43	1.13 $-$	0.34	4.81	1.38
$+/\approx/-$	0/5/1		0/0/6		0/0/6			

Symbols $-, \approx$ and $+$ represent that the competitor is worse than, similar to, and better than ICE according to the t-test at $\alpha = 0.05$.

B. Results of the Ablation Study

This subsection presents ablation study results for ICE across three complementary metrics of SR, CL and CR. Table B4, Table B5 and Fig. B2 show the experimental results of ICE compared with three variants under ablation settings across six GIRCP problems, based on 50 independent runs. These results provide deeper insights into how each algorithmic component contributes to different performance aspects.

Table B4 demonstrates the impact of each component on the reliability of conjecture generation. The variant with reduced pre-selection population (A1, $N_O = 1$) maintains excellent success rates comparable to ICE, achieving 100% on four problems and slightly lower rates on GIRCP2 (98%) and GIRCP5 (96%). This suggests that pre-selection, while beneficial, is not critical for basic algorithm reliability. In contrast, removing ADF encoding (A2) significantly reduces success rates to between 50% and 82%, demonstrating that ADF is essential for consistent conjecture discovery. Most dramatically, replacing the frequency-based DE operation with single-point crossover (A3) causes a catastrophic decline in success rates to between 12% and 20%. These results establish a clear hierarchy of component importance for reliability, with the frequency-based DE operation being most critical, followed by ADF encoding, while the specific pre-selection population size has a relatively minor impact.

The CL results in Table B5 reveal how each component affects the precision of discovered conjectures. Statistical testing shows that A1 performs comparably to ICE on five problems and worse on one (GIRCP5), indicating that reduced pre-selection primarily affects the quantity rather than quality of discovered conjectures. In stark contrast, both A2 and A3 perform significantly worse than ICE across all six problems. A2 achieves moderate CL values between 2.46 and 3.43, while A3 exhibits extremely low CL values between 1.08 and 1.21, indicating very low precision conjectures. Notably, while A1 shows standard deviations comparable to or lower than ICE, both A2 and A3 exhibit inconsistent precision levels. These results demonstrate that both ADF encoding and frequency-based DE operations are essential for discovering high-precision conjectures, with the latter having a particularly profound impact on conjecture quality.

Fig. B2 illustrates how each component affects convergence

behavior throughout the evolutionary process. ICE (red) consistently achieves superior convergence patterns, with progressively decreasing CR values across all problems. A1 (purple) shows convergence patterns similar to ICE in later generations for some benchmarks but exhibits greater variability, particularly in early generations. A2 (green) demonstrates moderate convergence in most problems but consistently underperforms ICE, especially in later generations. A3 (blue) shows the poorest convergence behavior, with highly unstable patterns and significantly higher CR values across all benchmarks. These patterns reveal that while reduced pre-selection primarily affects early convergence rates, removing ADF encoding or replacing frequency-based DE with single-point crossover fundamentally impairs the algorithm's ability to navigate the search space effectively throughout the evolutionary process.

The ablation results across all three metrics provide consistent evidence regarding the contribution of each algorithmic component. While all components contribute to ICE's performance, the frequency-based DE operation emerges as the most critical element, followed by ADF encoding, with the specific pre-selection population size having the least dramatic impact. These findings align with and reinforce the GR analysis presented in the main text, collectively establishing the significance of each component for effective mathematical conjecture discovery.