Innovative applications of genetic algorithms to problems in accelerator physics

Alicia Hofler,[1] Balša Terzić,[1,2] Matthew Kramer,[3] Anton Zvezdin,[4] Vasiliy Morozov,[1] Yves Roblin,[1] Fanglei Lin,[1] and Colin Jarvis[5]

[1]*Jefferson Lab, Newport News, Virginia 23606, USA*
[2]*Center for Accelerator Science, Old Dominion University, Norfolk, Virginia 23529, USA*
[3]*University of California, Berkeley, California 94720, USA*
[4]*Stony Brook University, Stony Brook, New York 11794, USA*
[5]*Macalester College, Saint Paul, Minnesota 55105, USA*

(Received 19 September 2012; published 9 January 2013)

The genetic algorithm (GA) is a powerful technique that implements the principles nature uses in biological evolution to optimize a multidimensional nonlinear problem. The GA works especially well for problems with a large number of local extrema, where traditional methods (such as conjugate gradient, steepest descent, and others) fail or, at best, underperform. The field of accelerator physics, among others, abounds with problems which lend themselves to optimization via GAs. In this paper, we report on the successful application of GAs in several problems related to the existing Continuous Electron Beam Accelerator Facility nuclear physics machine, the proposed Medium-energy Electron-Ion Collider at Jefferson Lab, and a radio frequency gun-based injector. These encouraging results are a step forward in optimizing accelerator design and provide an impetus for application of GAs to other problems in the field. To that end, we discuss the details of the GAs used, include a newly devised enhancement which leads to improved convergence to the optimum, and make recommendations for future GA developments and accelerator applications.

DOI: 10.1103/PhysRevSTAB.16.010101 PACS numbers: 07.05.Tp, 29.20.−c

I. INTRODUCTION

Accelerator physics deals with intricate systems which depend on many interrelated specifications/variables and physical quantities. One of its main goals is to design and operate accelerators so as to achieve an efficient interplay between these many quantities, thereby optimizing their performance. This is why genetic algorithms (GAs)—efficient, robust, multidimensional nonlinear optimization tools—are crucially important.

GAs have been used in the accelerator field as early as 1992 [1]. Initial use centered on the design and analysis of individual accelerator elements such as magnets and radio frequency (rf) cavities. With the successful optimization of the injector design for the Cornell energy recovery linac-based light source [2], interest in the applicability of GAs to larger accelerator design problems has intensified. Most applications are similar to the Cornell study where the goal is to find the optimal settings for the magnets and rf components in a beam line given a beam description and layout of the components [3–5] or additionally optimize the laser parameters for a photoinjector gun [6–8]. GAs have also been used to balance design and operating costs for a superconducting rf (SRF) linac [9] and to study alternative ring-based machine designs [10] for the International Linear Collider and in optics optimization studies for the Advanced Photon Source [11].

Machine element applications represent the earliest uses of GAs in the accelerator field. A free-electron laser application found the optimal order for the wiggler constituent magnets to minimize cumulative field error effects [1]. A superconducting magnet design tool developed at CERN produced a collection of magnet designs without meshing the superconducting coils and allowed the magnet designers to explore several possible designs subject to various engineering constraints [12,13]. GA-based tools have been used to design accelerator cavities meeting resonance and higher-order mode frequency requirements [14,15] and surface field constraints [16,17] and to diagnose a detuned SRF cavity installed in a multicavity cryounit [18].

Building on the precedent established with [2], in this paper, we expand the accelerator design problems to which GAs can be successfully applied. We demonstrate that GAs designed for multiobjective optimization are equally powerful when applied to single-objective problems, even problems that are difficult or computationally prohibitive to solve using standard nonlinear optimization techniques. For single-objective problems, GAs converge more quickly than standard nonlinear optimization techniques (when applicable) and are more efficient than systematic parameter scans. Our multiobjective optimization examples demonstrate how effective GAs are in searches where the interplay of the multiple dimensions

and objectives is unknown *a priori*. Each application presented, single- or multiobjective, is customized for a specific machine layout but serves as a roadmap for other machines. The GA methodology is general purpose, and to that end, we developed and use a GA framework that can be quickly and easily configured to work with different accelerator physics modeling codes. While in some cases presented here, such as betatron tune or dynamic aperture optimization, GA is essentially the only systematic approach, in other cases our framework extends the capability of existing codes by providing a straightforward interface to a powerful optimization routine greatly expanding the reach of GAs.

The remainder of the paper is outlined as follows. In Sec. II, we describe the theory of GAs. In Sec. III, we apply it to several problems in accelerator physics with a single objective to be optimized: search for the optimal working point in colliders, maximizing the dynamic aperture in a proton ring, and decoupling of a beam line. In Sec. IV, we apply GAs to problems that optimize multiple objectives: optimizing the dynamic aperture and momentum acceptance simultaneously for a proton ring and maximizing the brightness of an rf gun-based injector. Finally, in Sec. V we summarize the work presented, discuss its importance, and outline the possible future applications of GAs to other problems in accelerator physics.

II. BRIEF OVERVIEW OF THE THEORY OF GENETIC ALGORITHMS

Before we delve into explaining how GAs implement nonlinear optimization, a brief overview of the terminology is in order. The general statement of a minimization problem is

$$\begin{aligned} \text{minimize} \quad & f_i(x_1, x_2, \ldots, x_N) \quad i = 1, 2, \ldots, M, \\ & x_{\min}^{(j)} \le x_j \le x_{\max}^{(j)} \quad j = 1, 2, \ldots, N, \end{aligned} \tag{1}$$

where M is the number of objective (cost) functions to be simultaneously minimized and N is the dimensionality of the optimization problem—the number of independent variables varied. When $M = 1$, the minimization is a *single-objective* optimization, and $M > 1$ for *multiobjective* problems. $N > 1$ denotes a *multidimensional* optimization.

At the topmost level, a GA implements the principles of biological evolution to optimize a multidimensional nonlinear problem. Correspondence between evolution and multidimensional optimization is established if one views *genes* as independent variables and *individuals* as different sets of independent variable values and resulting objective function values. A group of individuals form a *population*, and successive populations are *generations*, the counterpart to iterations. As in multidimensional optimizations, GAs produce new values for the independent variables based on the characteristics of past individuals. The differences lie in how individuals are selected to create new ones and the creation methods themselves. Table I provides a summary of the parallel concepts in evolution (GA) and multidimensional nonlinear optimizations.

In general, *fitness*, a fundamental concept in GA optimizations, does not have a direct correlation in multidimensional optimization. In biological evolution, the strongest individuals in a population survive to produce offspring, and in GAs, fitness embodies this propensity for survival. A stronger or *fitter* individual is identified by its fitness value. Fitness is a function of the objective values specific to each GA and measures how well an individual meets the optimization goals. In its simplest form, for a single-objective optimization, an individual's fitness is its objective function value. In that case, evolution toward the fittest individual is clearly equivalent to the search for the optimal solution.

For multiobjective optimizations, the fitness function definition must account for multiple objective functions and accurately characterize the optimality of each objective value. This can be achieved with *dominance*. For a minimization problem—all objectives to be minimized—an individual A is said to dominate individual B if one or more of A's objective values is less than the corresponding values for B, and any remaining objective values are equal to B's. In the bounded-domain minimization [19],

$$\begin{aligned} f_1(x_1, x_2) &= x_1, \\ f_2(x_1, x_2) &= \frac{1 + x_2}{x_1}, \\ 0.1 \le x_1 &\le 1, \\ 0 \le x_2 &\le 5, \end{aligned} \tag{2}$$

shown in Fig. 1, individual A dominates individual B since both of A's objective values are less than B's. A is also said to be *nondominated* by B. A and B are both *feasible* because they are solutions to the objective functions, $f_1(x_1, x_2)$ and $f_2(x_1, x_2)$, and meet all specified constraints of the optimization, in this case, the bounded domain. In a single-objective optimization, the optimal solution is feasible and dominates all other feasible objective function values. Stated differently, the single-objective optimal

TABLE I. Correspondence between evolution and multidimensional optimization.

Evolution	Multidimensional optimization
Gene	Variable
Individual	Point in search space
Population	Set of points in search space
Mutation	Changing variable values
Recombination	Exchange of variable values between two points in search space
Generation	Iteration

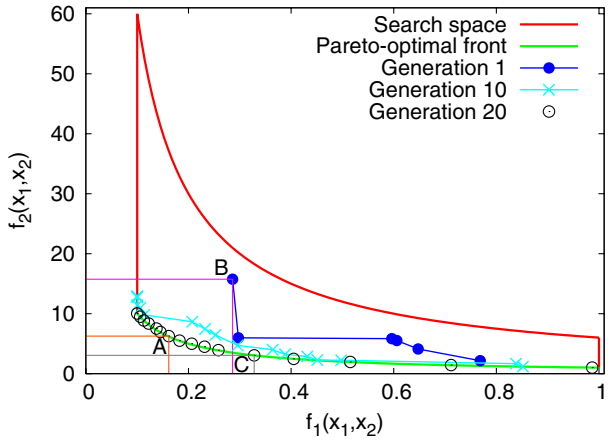

FIG. 1. The search space (red and green) and Pareto-optimal front (green) for the minimization of the system in Eq. (2) [19]. Pareto-optimal front estimates found by the Strength Pareto Evolutionary Algorithm (SPEA2) [20] after 1, 10, and 20 generations for 16 individuals are marked. The overlap between the search space and the rectangles with individuals A, B, and C at the top right vertices contain the points, if any, in the search space that dominate the respective individuals.

solution is nondominated by any other feasible objective function value. Since dominance, by definition, categorizes a set of objective function values, and optimal solutions are nondominated, it serves as a key criterion whereby multiobjective optimization with GAs is implemented.

Equation (2) illustrates another feature of multiobjective optimizations—conflicting objectives—that differentiate them from single-objective optimizations. In the example, the conflict is that minimizing $f_1(x_1, x_2)$ causes $f_2(x_1, x_2)$ to increase. When objectives conflict, the optimization has more than one equally valid solution, and these individuals form a *Pareto-optimal* front in search space. Each solution on this front is feasible, nondominated with respect to the other solutions on the front, and dominates at least one feasible individual in the search space. In Fig. 1, A and C are on the Pareto-optimal front. Although B and C are nondominated with respect to each other, B is not on the Pareto-optimal front since it is dominated by A. The task for a multiobjective optimization is to identify the Pareto-optimal front (a set of individuals), and dominance-based fitness functions are the best tools for this search.

Note that not all multiobjective optimizations have conflicting objectives. For instance, such a problem can be constructed from Eq. (2) if for the same bounded domain $f_1(x_1, x_2)$ is minimized, and $f_2(x_1, x_2)$ is maximized. The problem essentially reduces to a single-objective optimization to maximize $f_2(x_1, x_2)$. Its one solution at $(x_1, x_2) = (0.1, 5)$ gives $(f_1, f_2) = (0.1, 60)$ in Fig. 1 and forms a single-point Pareto-optimal front. Using a dominance-based fitness function, the problem can be solved either as a multiobjective or a single-objective optimization because the optimal solution to either shares the same characteristics: the solution dominates other feasible solutions and is nondominated.

The process a GA follows to identify the Pareto-optimal front for a multiobjective optimization is to first randomly generate a population of individuals to evenly sample the search space. The objective and fitness functions for each individual are evaluated. The GA then uses competition to select candidate individuals to form the *mating pool*. The fitness values of randomly chosen contestants from the generation are compared, and a copy of the contestant with the stronger fitness value is placed in the mating pool. Fitter individuals will win more competitions, have more copies of themselves in the mating pool, and therefore have a greater influence on the next generation. Pairs of individuals, *parents*, are taken from the mating pool to produce *offspring* pairs to populate the next generation. The offspring, modified copies of the parental gene pairs, are created through a process which simulates reproduction, whereby genes undergo: *recombination*—value exchange for the same independent variable—and *mutation*—single independent variable value adjustment. The value exchange in recombination can be a direct swap where the parent values selected for exchange are used without modification in the offspring. Alternatively, values used in the offspring can be functions of the selected parental values. Both recombination and mutation operations can be designed to enforce bounded-domain constraints. These operators also can be optionally switched off in an optimization, and whenever recombination is turned off, the GA method devolves to a Monte Carlo-based optimization. In a GA, the process of evaluating fitness and creating offspring is then repeated until desirable results are reached.

The choice of the population size should achieve a balance between the richness of genes and the speed of convergence. Too few individuals may not produce enough gene variety for successful search of a good optimum point. On the other hand, too many individuals may slow down the speed of convergence by increasing the number of function evaluations, thereby precluding the search from benefiting from the increased gene variety.

For the studies reported here, the Strength Pareto Evolutionary Algorithm 2 (SPEA2) [20] is used because it has produced good results in previous efforts at using GAs in accelerator design [2]. It is an *elitist* strategy. Because individuals are selected at random to participate in mating pool competitions, it is possible for some fitter individuals to be omitted from the mating pool for lack of being chosen to participate in a competition. Without intervention, these individuals and their influence on the optimization outcome are lost. An elitist strategy reserves the fitter individuals from each generation to supplement the set of individuals considered in subsequent generations to seed the mating pool, in essence, giving the optimization "memory." In SPEA2, reserved individuals are placed in

the *archive*, and only members of the archive are used to form the mating pool.

In each generation, the contents of the fixed-size archive are updated to contain the fitter individuals from the archive and the present generation. Fitness in SPEA2 tracks nondominance of individuals. For each individual in the archive and the present generation, SPEA2 tallies the number of individuals in the archive and present population that dominate the given individual. Under this definition, fitter individuals are nondominated and have tallies of 0. The archive, thus, contains at the end of each generation a cumulative estimate of the Pareto-optimal front. In Fig. 1, front estimates (nondominated individuals) for three generations are shown. As the generations proceed, the estimate of the Pareto-optimal front improves. While none of the individuals in the front for the first generation are in the final estimate, those individuals represent the best estimate for the Pareto-optimal front for the randomly generated individuals. Because GAs are population based, an advantage of GAs is that each generation contains an estimate of the Pareto-optimal front as evidenced by SPEA2.

A. General evolutionary algorithm code

We use automation systems that build on the Platform and Programming Language Interface for Search Algorithms (PISA) developed at ETH Zürich [21,22] and Alternate PISA (APISA) from Cornell University [2]. PISA is a modular test bed system for GAs. It separates the GA parent selection process from the optimization problem evaluation and population generation processes into two programs: the *selector* and the *variator*. This design easily allows different GA selection algorithms, selector programs, to be applied to several academic bounded-domain optimization problems for performance and convergence comparisons. APISA expands the available PISA problem types to include accelerator injector design with an interface to the beam dynamics simulation system A Space Charge Tracking Algorithm (ASTRA) [23]. It also provides support for strict inequality constraints. These constraints, unlike bounded-domain constraints, depend on the problem model evaluation to restrict the set of feasible individuals. Two examples for Eq. (2) are $f_2(x_1, x_2) < 20$ and $\sqrt{f_2(x_1, x_2)} < 10$. We developed a user-friendly, script-based hybrid of PISA and APISA and use it for all optimizations presented, except the rf gun optimization that builds directly on APISA [24–26].

As the name suggests, the selector is responsible for selecting individuals from among the population. The selector's tasks include calculating fitness values and selecting individuals for the mating pool and the archive. The selector program is complemented by the variator, whose job is to generate the offspring population. The variator is also responsible for calculating the value of each individual's objective function. For nontrivial physical problems, this is usually accomplished by dispatching a separate simulation. The variator dispatches the simulations with independent variables as unique input, and, upon their completion, parses their output to extract the values of the objective functions and inequality constraints. In the problems presented in this paper, the function evaluation is performed by sophisticated accelerator codes BEAMBEAM3D [27], ELEGANT [28], POISSON SUPERFISH [29], and ASTRA [23]. It is important to ensure the function evaluator reflects all important physics of the problem to be optimized, because the solution of the GA optimization is only as physical as the underlying model.

The simulations have been carried out on two large-scale computation facilities at Jefferson Lab: the high performance cluster [30], consisting of over 1500 cores, using a parallel Message Passing Interface paradigm and the batch farm cluster [31].

III. ACCELERATOR PHYSICS APPLICATIONS OF GENETIC ALGORITHMS: SINGLE-OBJECTIVE PROBLEMS

It often suffices to optimize only a single aspect of the accelerator performance by adjusting various variables of the design. The problem then reduces to a single-objective optimization. This section describes three such problems: (i) locating a near-optimal working point in a collider; (ii) maximizing the dynamic aperture in a collider ring; and (iii) decoupling of the beam optics in an injector. The first two problems arise in the design of the future electron-ion collider, while the last deals with an existing injector in Continuous Electron Beam Accelerator Facility (CEBAF) at Jefferson Lab.

While these problems address particular machines, the approach based on GAs presented here is general. GAs can be used to solve similar problems in other existing or future machines after accounting for different design layouts and parameters. While important and novel in their own right, these problems provide a template of how GAs can be used to optimize different aspects of accelerator design and performance.

A. Locating near-optimal working point in colliders

For over a decade, Jefferson Lab has been pursuing design studies for an electron-ion collider for future nuclear physics research, as outlined in the 2007 long range plan, DOE/NSF Nuclear Science Advisory Committee [32]. Based on CEBAF, the Medium-energy Electron-Ion Collider (MEIC) [33,34] would provide collisions between polarized electrons and polarized light ions or unpolarized heavy ions at multiple interaction points (IPs). The current plan adopts a staged path. An immediate goal of the electron-ion collider project is a low-to-medium-energy collider (MEIC) with center-of-mass energies up to 51 GeV. A future upgrade option is a high-energy collider with 100 GeV and greater center-of-mass energies. The

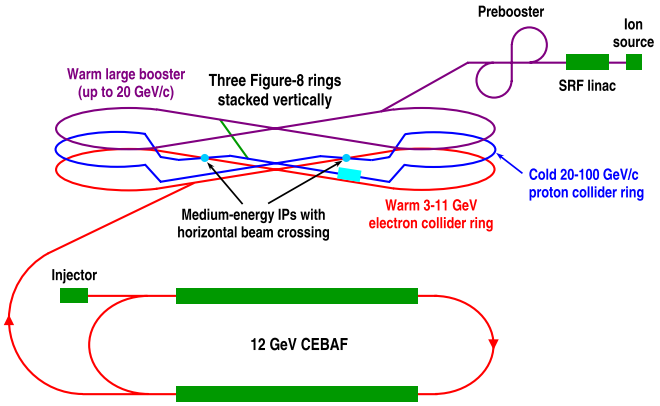

FIG. 2. Layout of the MEIC.

layout of the MEIC is shown in Fig. 2 and its model parameters in Table II.

For the purposes of the nuclear physics research, the two most important figures of merit in a collider are beam energies and collision luminosities. While different energy ranges may be better suited for different experiments, higher luminosity is preferred for the MEIC.

Luminosity and the long-term stability in a collider are sensitive to the synchrotron and betatron tunes of the two colliding beams. It is therefore imperative to select carefully the betatron and synchrotron tunes, also known as the *working point*, to assure the stable operation of the collider and achieve high luminosity. Here we apply a GA to locate a near-optimal working point for the MEIC design.

We use the term "near-optimal" to illustrate that there is no way of assuring that the GA will find the *global* extremum; instead, it finds a finite selection of "fit" points, from which the fittest is selected and pronounced near optimal.

TABLE II. Model parameters for the MEIC at Jefferson Lab [33,34].

Quantity	Unit	e^- beam	p beam
Energy	GeV	5	60
Collision frequency	MHz	750	
Particles per bunch	10^{10}	2.5	0.416
Beam current	A	3.0	0.5
Energy spread	10^{-3}	0.71	0.3
rms bunch length	mm	7.5	10
Horizontal bunch size at IP	μm	23.4	
Vertical bunch size at IP	μm	4.7	
Horizontal emittance (normalized)	μm	53.5	0.35
Vertical emittance (normalized)	μm	10.7	0.07
Horizontal β^*	cm	10	
Vertical β^*	cm	2	
Vertical beam-beam tune shift		0.029	0.0145
Damping time	turns	1516 (6.8 ms)	$\approx 2.4 \times 10^6$ ($\approx$ 11 000 ms)
Synchrotron tune		0.045	0.045
Ring length	m	1340.92	1340.41
Peak luminosity	$\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$	0.56×10^{34}	
Reduction (hourglass)		0.957	
Peak luminosity with hourglass effect	$\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$	0.54×10^{34}	

1. Optimization problem

At the heart of any GA is an objective function evaluator, which, given a set of independent variables, computes the objective function. The function evaluator here is a computer code which simulates beam-beam effects for a given working point and computes collider luminosity. The coherent beam-beam effects are dominated by the faster damping of the two beams, the electron beam. This study does not address beam-beam driven losses in the ion beam, which require simulation time scales to be on the order of the ion beam cooling time or longer. The goal here is to maximize the collider luminosity at the IP over the time scales on the order of electron beam damping time. Therefore, the objective function is the luminosity evaluated after twice the synchrotron radiation damping time (the luminosity is averaged over the last tenth of the damping time to avoid spurious results due to possible oscillations). The fact that this simulation covers only such a short initial period of collider's operation means that, in essence, it optimizes the *peak* luminosity.

In general, a simulation of beam-beam effects in a collider has two main components: tracking of particle collisions at IPs and transporting beams through the storage-collider rings. Colliding beams are modeled as bunches of macroparticles with the same mass-to-charge ratio. Each colliding beam bunch is divided in several computational slices, which affect each other through nonlinear kicks computed by the Poisson equation. The collision luminosity is calculated by summing the overlapping particle densities in each pair of interacting slices. At IPs, bunches of colliding beams interact by exchanging nonlinear kicks. The resulting nonlinear forces acting on particles are computed on a grid using standard particle-in-cell methods. In the present study we simulate the MEIC configuration with one IP, so that the transport of beams through storage-collider rings is modeled by one-turn linear maps. For more details on beam-beam simulations of the MEIC and the earlier design, see [35–37].

We simulate beam-beam effects in the MEIC with BEAMBEAM3D [27] simulation code, developed at Lawrence Berkeley National Laboratory. BEAMBEAM3D is a 3D, self-consistent beam-beam code which uses the shifted integrated Green's function method to solve the Poisson equation for electromagnetic fields on a grid. Each beam bunch imparts beam-beam kicks to the beam

bunch of the opposing beam with which it collides. The code is capable of running in both *strong-strong* mode, in which both colliding beams suffer perturbation by the beam-beam interactions in collisions, and *weak-strong* mode, in which only the ''weak'' colliding beam can be perturbed. BEAMBEAM3D code is parallelized so as to take full advantage of parallel computer architecture.

In the current implementation, we keep the synchrotron tunes fixed, and search the betatron tune space: x and y tunes for each beam, thus yielding a 4D problem. Therefore, this is a 4D, single-objective, nonlinear optimization problem.

This formalism can easily be extended to include also the synchrotron tunes, as well as the particle spin.

2. Restricting the search space

A systematic scan of the multidimensional tune space in search of an optimal working point is computationally prohibitively expensive. For example, covering each of the N betatron tunes with a modest resolution of 0.01 would require 10^{2N} function evaluations to cover the entire space; in our problem, we search over $N = 4$ betatron tunes only, which still results in staggering 10^8 multihour function evaluations. Without restricting the search space, even the GA implemented here would require much larger populations and many more generations to provide a reasonable working point, due to the vastness of the parameter space.

We restrict the search space to the most stable regions determined in the following way. Figure 3 shows a grid of both sum (denoted by black lines) and difference (denoted by green lines) resonances in the betatron tune space of up to order 7. The sum resonances always lead to dynamic instability due to resonant amplitude growth. Therefore, one generally wants to stay far away from them. The difference resonances, on the other hand, preserve a combination of the integrals of motion. They just cause exchange between the degrees of freedom, which, barring the situations such as beam envelope beating, keeps the motion bounded. Resonant lines are defined by $m\nu_x + k\nu_y = n$, where ν_x and ν_y are the betatron tunes, m, k, and n are integers, and n is the order of the resonance. The shaded regions are entirely devoid of sum resonances (black lines). It is also generally considered a bad idea to operate the collider with nearly integer tunes, which is why the regions near (0,0) and (1,1) are excluded from the search. The 16 regions (each with its mirror on the other side of the central point) cover only about 3.6% of the entire 2D tune space, which reduces the 4D search space and computational load by a factor of nearly 1000. With this realization, the search of the multidimensional parameter space becomes computationally tractable.

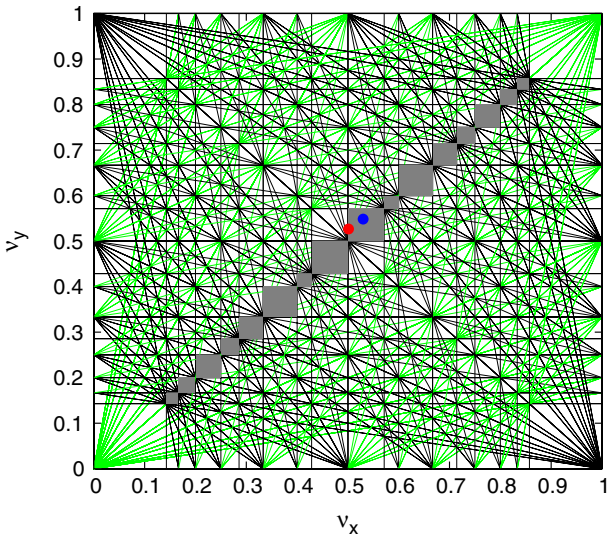

FIG. 3. A grid of sum (black) and difference (green) resonances up to 7th order. Shaded regions mark restricted search space for the GA, which is completely devoid of black resonance lines. The dots denote the optimal working point: red represents the betatron tunes for the proton beam, and blue for the electron beam.

3. Results

Given that each function evaluation may require hours of computing time on eight nodes of Jefferson Lab's computer cluster, it is imperative that the new algorithm locates a good working point within as few steps as possible. Each beam's tune can be located in any of the 16 regions of the tune space, which means that there is a total of $16^2 = 256$ areas of the tune space available for search. Randomly populating all of the 256 areas leads to solutions which slowly converge toward the near-optimal solution, in the sense that the most consequent generations are concentrated in the region in betatron tune space just beyond the half-integer resonance: $[0.5, 0.55]^2$, for each of the two colliding beams. The working point obtained in this comprehensive search exceeds design luminosity but is not optimal. To that end, we further restrict our search space to the single high-performing region [0.5,0.55] in each of the four tunes. This choice is also corroborated by the fact that PEP-II and KEK-B empirically converged to working points near the half-integer resonance. Figure 4 shows the luminosity for five generations consisting of 64 individuals which initially randomly sample $[0.5, 0.55]^4$ space. Within only 320 function evaluations, the algorithm located a working point at $(\nu_x, \nu_y) = (0.53, 0.548)$ for the electron beam and (0.501, 0.527) for the proton beam with luminosity of 7.05×10^{33} cm^{-2} s^{-1}, which exceeds design luminosity corrected for the hourglass effect of 5.42×10^{33} cm^{-2} s^{-1} by 30%. The enhancement of the collider's luminosity beyond the design value is due to the decrease in the beams' transverse size at the IP.

For the near-optimal working point, we compute the tunes of the subset of particles from each beam, and superimpose them on the resonance lines in Fig. 5. It is evident

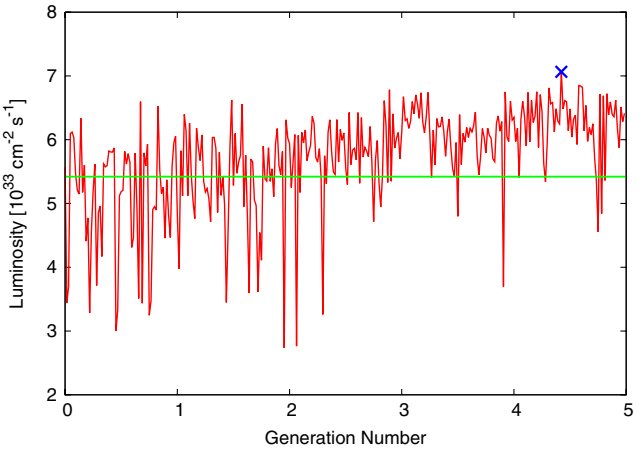

FIG. 4. GA at work: beam-beam simulation of the MEIC with five generations of 64 individuals each, sampling the 4D tune space $[0.5, 0.55]^4$. The green line represents the design luminosity. The optimization locates a near-optimal working point in fewer than 300 simulations (blue ×).

that tune footprints for both beams stay comfortably away from the unstable resonance lines. It is also interesting to note that the solutions obtained by the GA necessarily show favoring for the tune footprint with the proper orientation—away from the resonant lines. For all working points with low luminosity, one or more lower-order unstable resonance lines passes through the tune footprint.

The main factors limiting the closeness of the betatron tunes to the integer and half-integer resonances are their sensitivity to machine imperfections (alignment and field errors) and the particle loss due to single-particle scattering (Touschek and intrabeam). These are not a part of the physical model implemented here, which is why the near-optimal working point is so close to the half-integer resonance.

A more massive search, which includes more generations and individuals, can yield an even better working point. This is illustrated in the top panel of Fig. 6 which shows the improvement over design luminosity for each generation of a GA-based optimization where 128 individuals are evolved for 20 generations within the region [0.5,0.55] in each tune. This yielded the working point, $(\nu_x, \nu_y) = (0.525, 0.546)$ for the electron beam and (0.501, 0.501) for the proton beam, with luminosity of 7.53×10^{33} cm^{-2} s^{-1}, which exceeds design luminosity by 39%. However, it is not always obvious that additional computational work expended on a more detailed and longer search (here about 6 times) justifies the improvement in performance (here about 9%).

It is interesting to compare the results from a GA-based optimization with those from a traditional

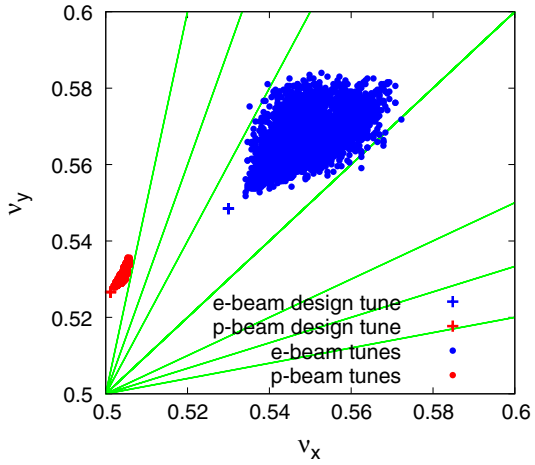

FIG. 5. Tune footprint for 4000 randomly selected representative particles from each beam for the near-optimal working point denoted by a blue × in Fig. 4, corresponding to $(\nu_x, \nu_y) = (0.53, 0.548)$ for the electron beam and (0.501, 0.527) for the proton beam. The electron beam is shown in blue, and the proton in red. The location of the near-optimal working point is denoted with two crosses, one for each tune.

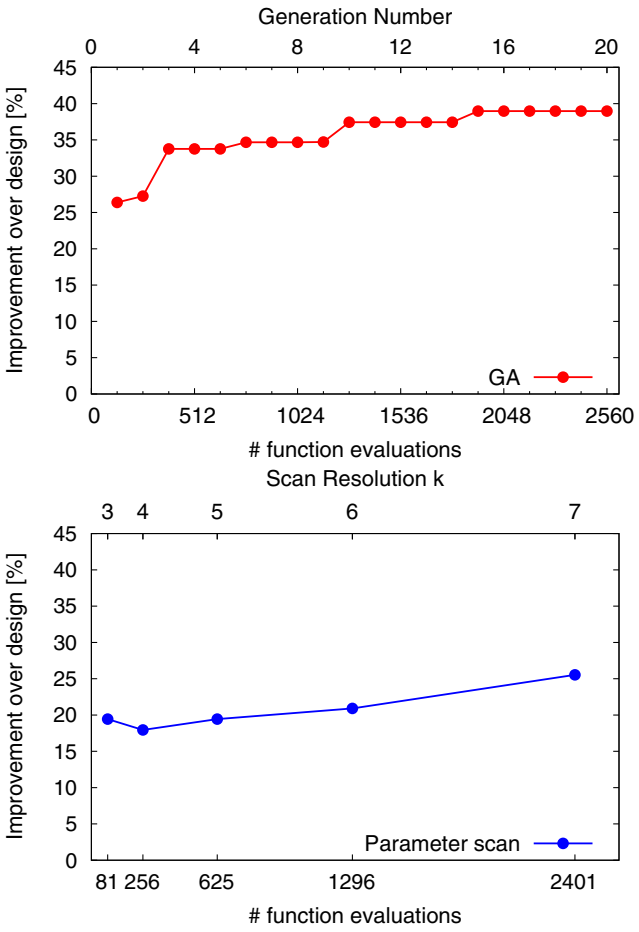

FIG. 6. Top panel: Improvement over the design luminosity after each generation for a GA-based optimization with 20 generations of 128 individuals in each. The improved working point is about 9% better than the one found in Fig. 4. Bottom panel: Improvement over design luminosity after a systematic parameter scan with resolution k in each parameter.

optimization such as a systematic parameter scan using the problem discussed in this section as an example. Note that in this instance, a systematic parameter scan requires k^4 function evaluations, where k is the number of evenly spaced discrete values for each of the four parameters (independent variables) and together uniformly sample the 4D search space. Extending to N dimensions, the number of function evaluations is k^N, and it is easy to see that this quickly becomes computationally prohibitive. In this $N = 4$ example, the ranges of parameters are quite small—[0.5, 0.55]—so the problem is marginally tractable by a (very coarse) parameter scan; this is not the case in the other applications presented here, which is why parameter scans are not implemented for those. Figure 6 compares the results obtained for the example case using GA-based optimization (top panel) and a systematic parameter scan (bottom panel). It is evident that the GA-based optimization is appreciably more powerful and efficient. More to the point, if one recalls that the first generation in a GA optimization (denoted by the leftmost red dot in the top panel of Fig. 6) randomly samples the entire allowable parameter space, then it is clear that the parameter scan even at the $k = 7$ resolution (the rightmost blue point in the bottom panel of Fig. 6) does not provide any improvement over such random sampling. (Note the curve for the parameter scan in Fig. 6 is not necessarily monotonic because as the resolution of the scan increases by 1, a different set of interior points is sampled; the proper refinement in resolution follows the sequence $k = 2, 3, 5, 9, \ldots$.)

4. Discussion

This study demonstrates that the GA is very efficient in finding the near-optimal working point for the collider. We recognize that the present physical model may be too simplistic to be used for the design of the real accelerator: it does not include nonlinear aspects of the collider rings, magnet imperfections, intrabeam scattering (IBS), the damping due to electron cooling of the ion beam, crab crossing by high integrated voltage SRF cavities, etc. These studies are currently under way. All the augmentations to the beam-beam simulations listed above will be implemented at the level of individual beam-beam simulations, and therefore confined to the function evaluator. However, the concept and implementation of the GA will remain intact. Therefore, this study serves as a proof of concept that GAs can efficiently optimize the collider working point.

Further sophistication of the GA-based beam-beam simulations will include the implementation of a multiobjective search in which additional objective functions will assure that the optimal working point is in a stable "neighborhood" in the tune space. Another important aspect of the collider that will be considered is its long-term operational stability. At the price of sacrificing the self-consistency of the physical model, the much-faster strong-weak simulations can enable the study of the medium- to long-term stability. This type of simulation will allow for optimization of the integrated luminosity, which is of greatest importance to the experiments.

B. Maximizing the dynamic aperture in a collider ring

In this section, we illustrate application of the GA to maximizing the dynamic aperture of a collider ring by optimizing its betatron tunes. Note that, to a large extent, the dynamic aperture optimization can be considered independently of the beam-beam interaction discussed in Sec. III A due to their different interaction ranges: the dynamic aperture is determined by dynamics of large-amplitude particles while the beam-beam interaction primarily affects the beam cores. Ultimately, of course, one would like to study these effects in combination.

To achieve the highest possible luminosity in a collider, the colliding beams should be focused to a small spot at the IP. Conceptually, the main challenge of designing a collider interaction region (IR) is compensation of chromatic effects associated with this strong focusing while preserving an adequately large dynamic aperture (a region in the transverse plane of stable particle motion).

One approach to IR design is presented in [38–40]. It involves installation of a dedicated chromaticity compensation block (CCB) between a beam extension section (BES) and a final focusing block (FFB) (see Fig. 7). In the CCB, certain symmetries of the beam orbital motion and dispersion are created using a symmetric arrangement of dipoles, quadrupoles, and sextupoles [40]. Namely, the particle's horizontal and vertical betatron trajectory components must be either symmetric or antisymmetric with respect to the center of the CCB,

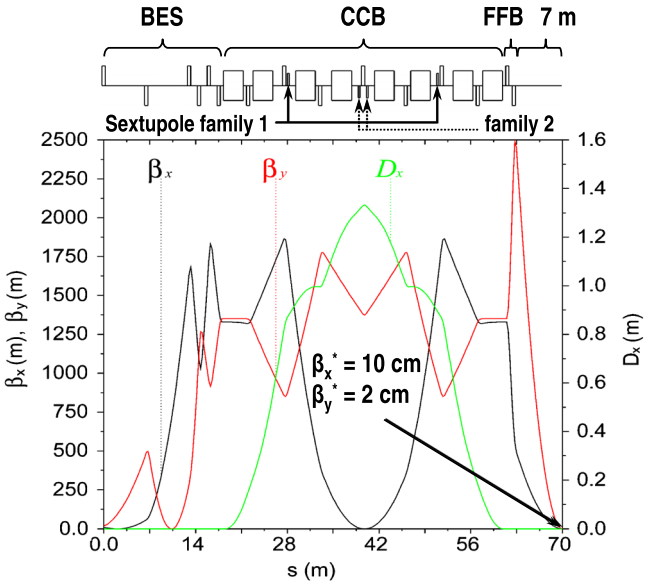

FIG. 7. Linear optics of the ion ring's IR.

and the quadrupole field component must be symmetric, while the dispersion and the sextupole field component must have symmetries opposite to that of the horizontal trajectory component. These symmetries then allow simultaneous compensation of the 1st-order chromaticities and chromatic beam smear at the IP without inducing significant 2nd-order aberrations and therefore helping preserve the ring's dynamic aperture.

The above IR design concept is implemented [40–42] in the prototype electron and ion collider ring lattices of the MEIC [34]. As shown in Fig. 2, the collider rings have geometrically matching figure-8 shapes consisting of two 120° arcs connected by two straight sections crossing each other in the middle at 60°. Each collider ring contains two IRs located in the two straights and is arranged in a twofold symmetric way. Figure 7 shows the optics of the first half of the ion IR from the start of the beam extension to the IP, with the second half of the IR being mirror symmetric around the IP. The parts of the straights not taken up by the IRs are filled with dispersion-free focusing-drift-defocusing-drift (FODO) cells. The rings' betatron tunes are adjusted by changing the betatron phase advance in the straights' FODO regions. The ring optics outside of these regions is maintained undisturbed by using a few quadrupoles at the ends of each FODO region to match every new FODO setting to the fixed optics at the regions' ends. The two straight FODO regions of each ring are identical and are adjusted in the same way simultaneously so that the ring's twofold symmetry is preserved.

Since the designs of the electron and ion collider rings are similar, below we will focus on the ion ring. This is a rather challenging case due to the ring's relatively large horizontal $\xi_x \equiv \partial\nu_x/\partial\delta$ and vertical ξ_y linear chromaticities of -319.7 and -397.0, respectively. Contributions of the IRs to ξ_x and ξ_y are -296.0 (92.6%) and -375.0 (94.5%), respectively. The IR design shown in Fig. 7 satisfies the symmetry requirements discussed above. Two sextupole families arranged symmetrically in each CCB as shown in Fig. 7, provide simultaneous compensation of both the horizontal and vertical linear chromaticities. For the nominal (ν_x, ν_y) working point of (0.31, 0.32), this chromaticity compensation approach results in an excellent momentum acceptance of $>14\sigma_\delta$.

The factors limiting this compensation technique include a small finite angular spread in the beam even after expansion, violation of the symmetry conditions outside of the CCB, and higher-order effects, such as amplitude-dependent tune shift and higher-order chromaticities. Therefore, further optimization of the nonlinear dynamics may be required using additional sextupole and octupole families. Another important optimization aspect is the choice of the betatron tunes. Even though, as described above, the phase advance changes only in the FODO regions of the straights, while most of the contribution to the nonlinear effects comes from the IR's, the choice of the betatron tunes sets their position with respect to beam resonances and determines the phase advance between the two IP's and between the consecutive passes through the same IP, which can enhance or suppress certain chromatic and geometric perturbations.

1. Optimization problem

In case of the MEIC, the momentum acceptance is already adequate just after the linear chromaticity compensation. Therefore, we first focus on optimization of the dynamic aperture. There are a number of techniques available for this task such as optimization of the resonance driving terms [43,44], minimization of the tune diffusion rate [45–47], and direct maximization of the dynamic aperture area [48,49]. The GA is not specific to any of these techniques and can be applied using any one or a combination of those. Below we illustrate optimization of the betatron tunes by maximizing the area of the dynamic aperture.

The MEIC ion collider ring is simulated using ELEGANT [28] particle tracking code. All ring components are modeled as canonical kick elements with exact Hamiltonians retaining all orders in momentum offset. The magnet fields are approximated as hard edge, and the lattice is assumed ideal, i.e., containing no errors. Betatron tune adjustment and appropriate rematching to preserve the ring optics outside of the straights' FODO cells, as discussed above, are automated in ELEGANT. Also, in each case, both the horizontal and vertical linear chromaticities are compensated down to zero using the two sextupole families, as described above.

The dynamic aperture of the ion collider ring is obtained for each betatron tune setting using ELEGANT in the following way. An on-momentum particle is launched parallel to the beam axis at the entrance into one of the CCB's. It is then tracked for 100 turns using kick-based 2nd-order symplectic integration. The particle initial coordinates are gradually moved away from the beam center along 13 rays originating at the center and spaced out equally in the upper half plane. The distance of the launch point from the center is increased along each ray until the particle stability is lost within 100 turns. The stability border is determined to within about ± 0.1 mm precision. The edge-of-stability points located on different rays are connected by straight lines outlining a stability region in the transverse plane, or, in other words, the dynamic aperture. The area of this region is used as the objective function. The choice of a small number of turns for finding the dynamic aperture is driven by the computational time considerations and is justified by the fact that our main goal is to investigate the relative effect of different betatron tunes rather than to study the ring's long-term stability.

Figure 8 shows the dynamic aperture after the linear chromaticity compensation for the initially chosen working point of (0.31, 0.32). These (ν_x, ν_y) values seemed to be a reasonable initial guess based on their location on the beam

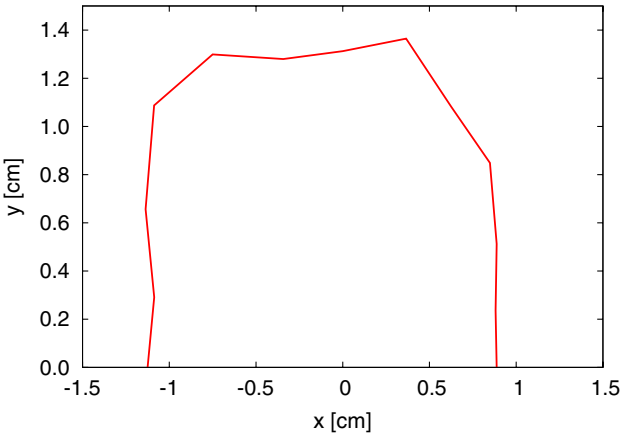

FIG. 8. Dynamic aperture after the linear chromaticity compensation for the initially chosen working point of $(\nu_x, \nu_y) = (0.31, 0.32)$.

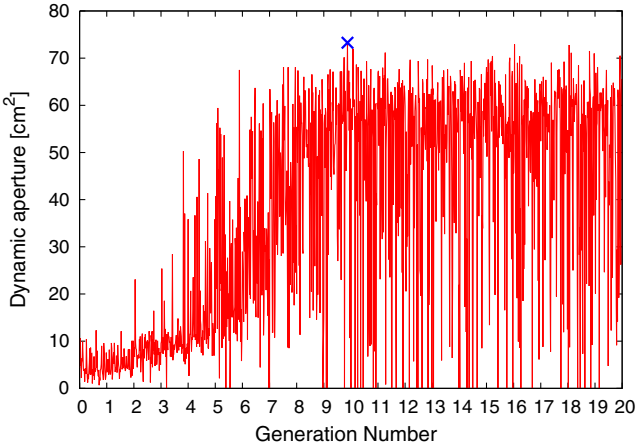

FIG. 9. Maximization of the dynamic aperture for the MEIC ion collider ring: 64 individuals sampling the $[0, 1]^2$ domain, evolved using GA over 20 generations. The blue $\times$ denotes the maximum value of the dynamic aperture.

resonance grid and the running experience of the existing machines. The dynamic aperture in Fig. 8 is reasonably large, especially considering the large compensated values of the natural chromaticities. However, due to the large beam extension required to achieve the ambitiously small IP β values at the MEIC, the horizontal and vertical sizes of the dynamic aperture correspond to only $\sim 4\sigma_x$ and $\sim 15\sigma_y$, respectively. Since a practical design typically requires a dynamic aperture above 6σ in both transverse dimensions with all imperfections and realistic effects taken into account, further optimization of the dynamic aperture even within this simplified model is required.

We combine the GA described in Sec. II with ELEGANT code [28] as a function evaluator. The independent variables are the two fractional parts of the betatron tunes, ν_x and ν_y, while the objective function is the dynamic aperture, evaluated as described above. The domain of the two independent variables is 0 to 1. Therefore, this is a 2D, single-objective, nonlinear optimization problem.

2. Results

Figure 9 shows the results of the GA run with 64 individuals and 20 generations, sampling the entire $[0, 1]^2$ domain in fractional betatron tunes. The optimal value of the dynamic aperture is found in the tenth generation (after about 640 function evaluations). By the sixth generation, the optimal value becomes prevalent, signaling that the algorithm has converged on the location of the optimal fractional betatron tunes. Figure 10 plots the dynamic aperture data from Fig. 9 as a function of the fractional betatron tunes. From the clustering of the points near the (1, 0) point in Fig. 10, it is evident that the GA has converged to this region, having identified it as the one with optimal fractional betatron tunes. The largest dynamic aperture occurs at a working point of $(\nu_x, \nu_y) = (0.994, 0.001)$. Precipitous drops in the area of the dynamic aperture in Fig. 10 correspond to the points too close to the (1, 0) integer resonance.

The choice of the optimal working point for the beam-beam interaction described in Sec. III A and for the maximum dynamic aperture discussed in this section are determined by different processes. Nevertheless, in comparing the two cases, one has to consider that the ring model used in this section includes two IPs, or two superperiods; therefore, all phase advances are doubled in comparison to Sec. III A, which only assumed one IP. The two cases then demonstrate similar optimal working points close to half-integer resonances. This is a general feature of linear or properly linearized systems; the most stable dynamics in these systems are near half-integer resonances because they are devoid of other low-order resonances (of order ≈ 2 and above). This result confirms that the model's chromaticity compensation scheme successfully suppresses the nonlinearity introduced by the final focus. Selecting a working point so close to half-integer resonances is perhaps not realistic, especially for ion beams that do not have synchrotron damping. To find a more realistic solution, the model has to be extended to include magnet errors, beam-beam interaction, IBS, etc. In fact, in a practical design optimization, one has to

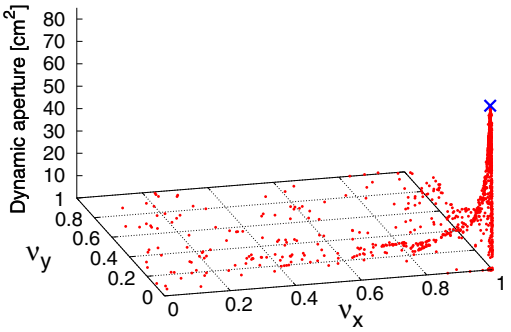

FIG. 10. Dynamic aperture versus the fractional betatron tunes for the simulation in Fig. 9.

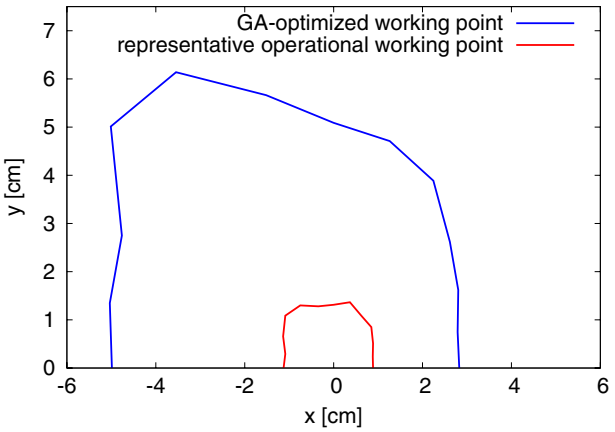

FIG. 11. Comparison of the dynamic aperture for the initial $(\nu_x, \nu_y) = (0.31, 0.32)$ working point (red line) to that for the optimal (0.994, 0.001) working point marked with the blue $\times$'s in Figs. 9 and 10 (blue line).

maximize the correctable dynamic aperture over an ensemble of machines with randomly generated realistic magnet alignment and field errors. However, the same algorithm as in this conceptual demonstration can be applied and offers the same advantages over an optimizer that searches for a local extremum as discussed earlier.

Another noteworthy observation is that, due to the low dimensionality of this optimization problem, we are able to search the whole tune space. However, the optimal working point still falls into one of the stability regions discussed in Sec. III A, which validates the procedure for the search domain reduction used in that section.

Figure 11 illustrates the improvement to the dynamic aperture attained by optimizing the working point: the initial dynamic aperture from Fig. 8 is compared to the largest-area dynamic aperture obtained in the optimization, corresponding to the blue $\times$'s in Figs. 9 and 10.

In comparison, a systematic scan of the entire 2D fractional betatron tune space, at a resolution of 0.01 would require $100^2 = 10\,000$ function evaluations. By employing a GA in our search, we reduced the required number of function evaluations to a few hundred—a computational savings of at least 1 order of magnitude.

C. Decoupling of the beam optics in the injector

The Continuous Electron Beam Accelerator Facility (CEBAF) is a superconducting facility located at Jefferson Lab. It provides a continuous electron beam of up to 6 GeV for use for nuclear physics experiments in up to three experimental halls simultaneously. (See the layout in Fig. 12.)

The beam is generated at the electron gun equipped with a GaAs photocathode. A circularly polarized laser beam impinges on this cathode and allows for polarized electrons to be produced with a longitudinal polarization in excess of 85% and currents as high as 200 μA.

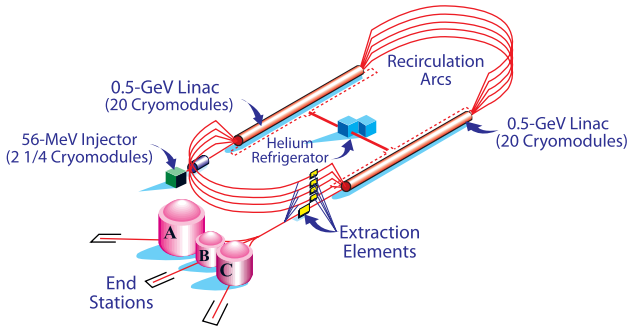

FIG. 12. Schematic of the CEBAF accelerator.

The electrons have an initial kinetic energy of 130 keV, and are then accelerated, bunched, and compressed in the injector to an energy of a few tens of MeV depending on the linacs' energy gains and the desired energies in the experimental halls. After the injector, acceleration to the experiment energies is achieved in two superconducting linacs, set in an antiparallel configuration and connected by asynchronous recirculation arcs. CEBAF can thus deliver beams between 0.6 and 6 GeV. The delivery of very high-quality beams with energy spread less than 3–4 $\times 10^{-5}$ and geometric emittances on the order of 10^{-9} mrad is routinely achieved.

Parity-violating experiments are the most challenging in terms of beam quality amongst the wide range of experiments performed at CEBAF. They require the helicity-correlated beam positions and angles to be controlled at the level of nanometers/nanoradians. Most of these correlations originate at the laser table and, while they can be controlled, they cannot be completely eliminated. To reach the desired level of accuracy, one has to rely on the natural adiabatic damping occurring during acceleration to reduce the helicity-correlated position and angle differences to the desired tolerances. One of the implications to the beam transport is that one has to suppress the transverse coupling in the injector proper to prevent projected emittance growth from occurring [50].

Coupling in the machine originates mainly from the SRF cavities. Skew quadrupoles are installed in the linacs between the rf zones to compensate for this. The injector is instrumented with eight skew quadrupoles located along the beam line where the SRF cavities are installed. The most damaging emittance growth potentially occurs in the injector proper since the beam is accelerated from 130 keV at the cathode to 60 MeV for a standard one-pass parity experiment. Here we describe the scheme we have adopted for correcting such transverse coupling in the injector transport line.

To recover from the emittance growth that occurs because of the x-y coupling, one has to alter the strengths of both the skew and normal quadrupoles. One approach is to measure the 4×4 transport matrix across the SRF modules and then use the skew quadrupoles to decouple the system and normal quadrupoles to rematch the beam line [51].

An alternative method which lends itself well to real-time corrections is to instead measure the beam shape with beam profile monitors. For this test, we use wire scanners equipped with x, y and 45-degree wires. By using four such scanners placed at the proper phase advances, it is possible to access all the off-diagonal terms of the transport matrix. The algorithm then simply becomes one of minimizing all these cross terms by making use of the available skew quadrupoles. For the sake of demonstration, we simulate a simple system where one gets the off-diagonal terms from a model and iterates the GA until they are nulled. This can also be solved with traditional nonlinear least square optimization. The GA starts being better suited to this kind of problem when one also includes a number of other constraints on the optics such as maximum beam size and phase advance constraints. Other groups have started using ELEGANT together with GAs for lattice design and optimization [52].

1. Optimization problem

We again combine the GA described in Sec. II with ELEGANT code [28] as a function evaluator. The independent variables are strengths of the six skew quadrupoles, allowed to vary within their entire operational ranges. The objective function value computed by ELEGANT simulation is the sum of squares of the cross terms at four different locations in the beam line with proper phase advances, $S = \sum_{i=1}^{4} (\sigma_{xy}^{(i)})^2$, and should, therefore, be minimized. Hence, this is a 6D, single-objective, nonlinear minimization problem.

2. Heuristic shrinking of the search space

The GA simulation of the beam decoupling proved to be rather well behaved: as the objective function decreases, individuals converge to a smaller region of an area of search space previously occupied. We observe that the individuals neither converge to multiple disjoint subareas nor do they drift away from the area where they initially settle (Fig. 13). Moreover, after many iterations within some search area, the speed of convergence starts to decrease substantially, as shown by the blue line in Fig. 14. However, if the search space is restricted to contain only its small subarea where the individuals converge, the rate of convergence of the algorithm drastically increases, as shown by the red line in Fig. 14. This happens because arriving close to a local optimum requires high precision gene variation (achieved through mutation in small steps). Since mutation amplitude depends on the search area size, decreasing the area decreases the average mutation amplitude, which, in turn, increases the speed of convergence.

We improve the rate of convergence of the GA by enhancing it with an area-shrinking algorithm which reduces the search area as soon as individuals converge within some fraction of the original search space. The

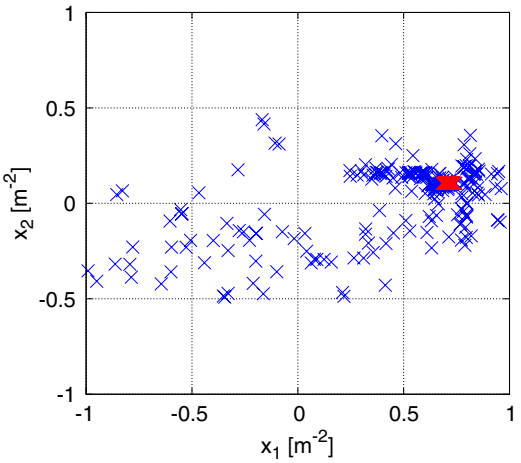

FIG. 13. Spread of all the individuals in the simulation in the (x_1, x_2) subdimension of the entire search space (blue points), and those with the value of the objective functions below 10^{-10} (red). The simulation shows 25 generations with 24 individuals apiece.

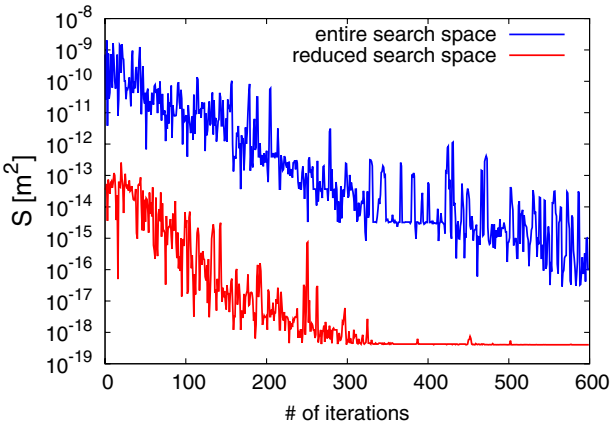

FIG. 14. Simulations over the entire search space (blue) and the reduced search space (red).

shrinking is done in each search variable independently, quantified by two parameters: shrinking threshold and margin size. If the total size of the search space in each dimension becomes smaller than some fraction of the original search space denoted by the shrink threshold, the search space shrinks. The new search space then becomes the range between these two points plus the extra space on each size determined as the margin size parameter times the range size.

3. Results

The simulations show that enhancing the GA with the shrinking algorithm noticeably improves the speed of convergence, on average by about 15%. Figure 15 shows a representative comparison between optimizations using GA without and with shrinking enhancement.

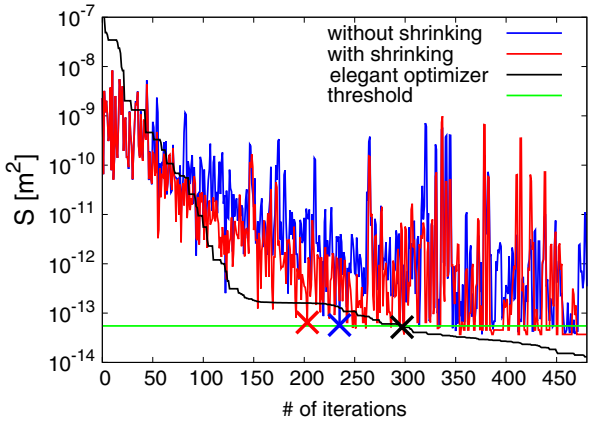

FIG. 15. A representative comparison between the GA simulation of 20 generations and 24 individuals without shrinking enhancement (blue) and with (red). Also plotted is the simulation with the ELEGANT simplex optimizer (black). The number of iterations needed for each method to converge to within the predefined threshold S is marked with an $\times$ of corresponding color.

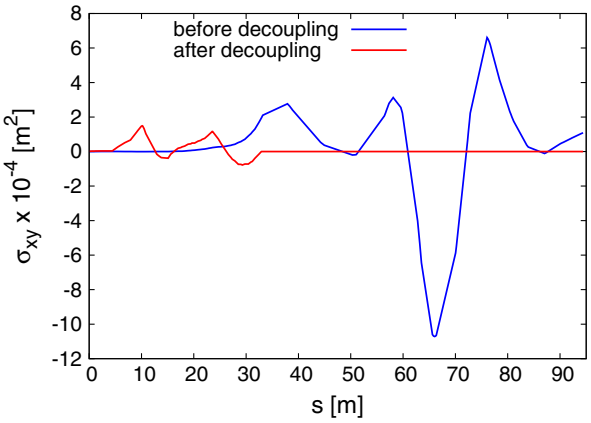

FIG. 16. σ_{xy} correlations at the end of the injector line before and after the decoupling procedure. Not shown are the other off-diagonal terms that are nulled as well.

The purpose of this optimization is not arriving at the global minimum, but rather obtaining an acceptable local minimum, $S = 5.5 \times 10^{-14}$, reasonably quickly. For each set of simulation parameters (shrinking threshold, margin size, η_{mut}, and η_{rec}), several simulations, each with a different random seed for the initial distribution of individuals in the search space, are executed and then averaged. The mutation and recombination factors, η_{mut} and η_{rec}, are described in the Appendix.

The fastest convergence below the prescribed S threshold of 5.5×10^{-14} is achieved with $\eta_{\text{mut}} = 10^6$, $\eta_{\text{rec}} = 0$, shrinking threshold of 0.99, and margin size of 0.001. These simulations take about 261 ± 35 steps, depending on the initial random seed. The same simulations without shrinking takes about 302 ± 125 steps.

In comparison, the simplex optimizer in ELEGANT reaches its final local minimum in about 300 iterations. (Other ELEGANT optimizers, such as grid-based and random walk, did not reach the same level of performance.) It is important to remember that the GA is *globally* convergent, unlike the simplex optimizer in ELEGANT, which finds the *local* extremum nearest to the initial starting point of the search. Because of the nature of the algorithms, the ELEGANT simplex optimizer only searches for a local minimum and is therefore sensitive to the starting conditions and the parameters of the algorithms. Achieving the optimal number of iterations requires significant hand tuning, and in some instances fails altogether. Therefore, the GA is both more efficient and more robust. The resulting decoupling of the injector line is illustrated in Fig. 16.

This study demonstrates the viability of GAs for an electron beam decoupling. The algorithm shows robust convergence to near-optimal points. An important improvement to automatically reduce the area of search produces a measurable improvement of 15% in algorithm efficiency. This approach may also prove useful in implementation of GA for other problems.

IV. ACCELERATOR PHYSICS APPLICATIONS OF GENETIC ALGORITHMS: MULTIOBJECTIVE PROBLEMS

Multiobjective optimization problems in accelerator physics were tackled in two ways until the advent of GA-based frameworks. The first was to simplify the problem as much as possible to reduce it to a single-objective optimization or to construct a weighted-sum single-objective function from the multiple objectives and then apply standard single-objective optimization techniques to find a solution. The downside to this method is that creating a single objective for a multiobjective problem can result in an objective function whose behavior does not accurately reflect the dynamics of the multiobjective system. Further, the optimal solution of the ersatz problem may not produce the desired results in the multiobjective system. Weighted-sum objective functions are unsatisfactory because the solution found and the rate of convergence depend heavily on the weighting factors. Tuning the weighting factors can be frustrating and unfruitful. The second approach was to perform parameter scans on a subset of the objectives and analyze the results. Systematic parameter scans are often prohibitively computationally intensive, especially for problems with more than say 2–3 degrees of freedom (e.g., see III A 2). While these approaches are effective in spite of the pitfalls, ensuring the solutions found are robust and not just optimal in a local sense is onerous and often impossible. These techniques were appropriate in computationally limited resource environments. With ever increasing computational power at decreasing cost, multiobjective optimization problems can now be addressed without the compromises of the past.

GA-based algorithms are adept at both single- and multiobjective optimization. In contrast to the first approach which involved constructing a single-objective problem description for the multiobjective system, the complete multiobjective optimization can be solved with GA methods. This leads to greater confidence in the solutions found because fewer simplifying assumptions are required. GA-based algorithms, also, can quickly identify promising regions in the global search space, a marked improvement over parameter scans.

In this section, we present two GA-based multiobjective optimizations. First, we reconsider the single-objective dynamic aperture optimization (III B) seeking to additionally minimize chromatic effects. The second optimizes the brightness of an rf gun-based injector and uses constraints based on model evaluator results to restrict the search. Both problems are sufficiently complex that parameter scans are untenable, yet they are tractable with GA methods.

A. Optimization of the dynamic aperture and chromaticity correction in a collider ring

Section III B discusses challenges associated with organizing a low-β IP in a collider and describes an approach to mitigate them using field and orbital symmetries in the IR design [41]. This approach is applied to the challenging case of a high-chromaticity small-β MEIC ion collider ring [40,42]. In Sec. III B 1, the GA is used to maximize the dynamic aperture by finding the optimal fractional betatron tunes ν_x and ν_y in the range between 0 and 1. The optimum (ν_x, ν_y) working point is found to be (0.994, 0.001). However, that single-objective optimization does not consider the impact of varying the betatron tunes on the ring's momentum acceptance. After compensating the linear chromaticities, the momentum acceptance is determined by the higher-order ones, which depend on the choice of the working point. In general, the smaller the 2nd-order chromaticities, the larger the momentum acceptance. Large 2nd-and higher-order chromaticities drive particles into beam resonances causing their loss. This necessitates an approach in which the GA is used to optimize dynamic aperture and momentum acceptance *simultaneously.*

1. Optimization problem

We use ELEGANT to compute two objective functions that are to be optimized simultaneously, namely, the dynamic aperture and the 2nd-order chromatic function $\xi^{(2)}$ defined as a sum of relative magnitudes of the 2nd-order chromaticities: $\xi^{(2)} \equiv ||\partial^2\nu_x/\partial\delta^2| - 1000| + ||\partial^2\nu_y/\partial\delta^2| - 2500|$. The dynamic aperture is obtained following the procedure described in Sec. III B 1. The 2nd-order chromaticities are calculated in ELEGANT by concatenating the ring's transfer matrix for a set of $\Delta p/p$ values and finding the trace of the off-momentum matrices [28]. We then use the GA described in Sec. II to optimize these two objective functions simultaneously. In this case, optimization entails minimizing the inverse of the dynamic aperture and the value of the chromatic function $\xi^{(2)}$. There are two independent variables, as before: the fractional parts of the betatron tunes, ν_x and ν_y, varying between 0 and 1. Therefore, this is a 2D, multiobjective, nonlinear optimization problem.

2. Results

Figure 17 shows the results of the GA run by plotting the Pareto-optimal front of the chromatic function $\xi^{(2)}$ versus the inverse dynamic aperture $1/A$ after 24 generations of 64 individuals. Note the resemblance of Fig. 17 to Fig. 1 resulting from the conceptual similarity of the underlying problems. Two conclusions can be drawn immediately from Fig. 17. First, the Pareto front is composed of a few sections that form individual islands in the (ν_x, ν_y) tune space. Since the islands are isolated from each other, locating the global optimum would have been impossible using conventional optimization techniques, which would only converge to the nearest local extremum. Second, there is an inverse relationship between $\xi^{(2)}$ and $1/A$, i.e., the objectives conflict with each other. This means the point with the largest dynamic aperture (point A in Fig. 17 with the minimum $1/A$ value) has the largest $\xi^{(2)}$ and, therefore, the smallest momentum acceptance and vice versa (point C in Fig. 17).

From the nonlinear dynamics point of view, the choice of the optimal working point is a balance between the momentum acceptance and dynamic aperture being both reasonably large such as point B in Fig. 17. Note that, in case of MEIC, due to its aggressively small collision point β values, further nonlinear optimization is required to have both the dynamic aperture and momentum acceptance adequately large [40]. This can be done by optimizing multiple sextupole and octupole families. The argument for choosing an optimal working point that gives a balance

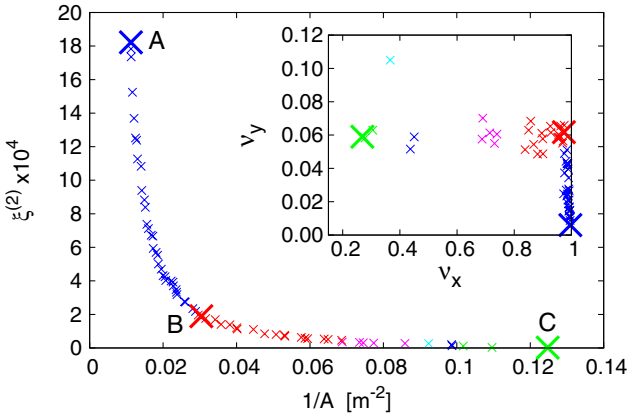

FIG. 17. Pareto front after 24 generations of 64 individuals. The large ×'s denote representative points A, B, and C.

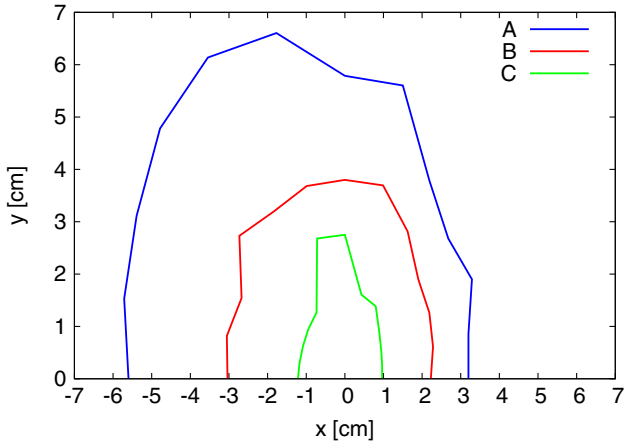

FIG. 18. Dynamic aperture plots corresponding to points A, B, and C in Fig. 17, as indicated by the color.

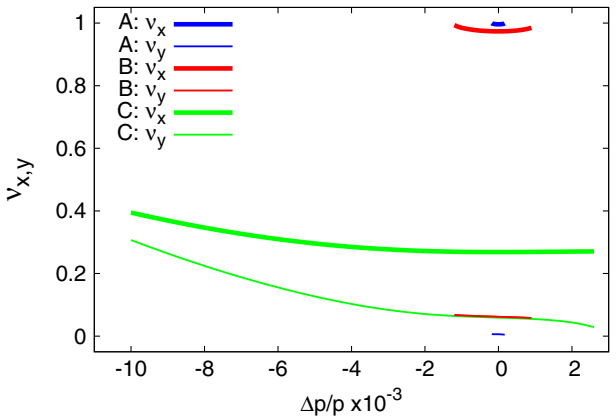

FIG. 19. Fractional betatron tunes ν_x and ν_y as a function of momentum offset $\Delta p/p$ for points A, B, and C in Fig. 17, as indicated by the color.

between the dynamic aperture and momentum acceptance, however, remains valid.

The dynamic aperture and momentum acceptance corresponding to points A, B, and C are compared in Figs. 18 and 19, respectively. Figure 18 plots the dynamic aperture in the x-y space. Figure 19 shows the fractional betatron tunes ν_x and ν_y as a function of momentum offset $\Delta p/p$. The horizontal extent of the lines in Fig. 19 indicates the size of the momentum acceptance.

B. rf gun optimization for injector brightness

As demonstrated in III C, for linac-based accelerators, the beam quality at the exit of the particle source or injector determines the quality of the final beam of the entire machine. Another example is light source brilliance and 6D brightness [53],

$$B_n = \frac{N}{\varepsilon_{n,x}\varepsilon_{n,y}\varepsilon_{n,z}},$$

where N is the number of electrons in the particle bunch and $\varepsilon_{n,x}$, $\varepsilon_{n,y}$, and $\varepsilon_{n,z}$ are the normalized transverse and longitudinal emittances of the injector. B_n directly affects the brilliance of a linac light source. In general, it is crucial for the injector to produce the highest quality beam to ensure that each new machine meets ever more aggressive application requirements, and for a linac-based light source, this translates to maximizing B_n. For a fixed bunch charge, minimizing the injector $\varepsilon_{n,x}$, $\varepsilon_{n,y}$, and $\varepsilon_{n,z}$ maximizes B_n.

Here we describe an optimization system that can vary the shape of the field profile of an rf gun in response to the performance of the beam dynamics [24–26,54]. Often rf guns are used in linac-based light sources. The typical rf gun design consists of a half-cell cavity containing the photocathode optionally joined to one or more full cells. The photocathode is inside the half-cell cavity at the center of the upstream cavity end plate (centered at the origin in Fig. 20). The peak field is at the cathode, and for multicell designs, the field profile is balanced, meaning the peak field amplitude is equal in each cell as shown in Fig. 21. We use the GA optimization to investigate if this design can be improved to increase source B_n. We discuss the field generation method and present results for an rf gun-based injector similar to the rf gun injector developed by the Photo Injector Test Facility Zeuthen (PITZ) [55]. The PITZ gun is the injector for the Deutsches Elektronen-Synchrotron (DESY) Free-Electron Laser in Hamburg (FLASH) [56]. Note that, while the optimization system is used to minimize normalized emittances in this example, it is general purpose and can be used to optimize rf gun-based injectors with respect to other beam dynamics and cavity performance criteria. Also, in contrast to varying an idealized numerical model of the field profile [24] or interpolating between field profiles in a catalog of externally produced profiles for different gun geometries [57], physical dimensions of the gun are varied and a field solver

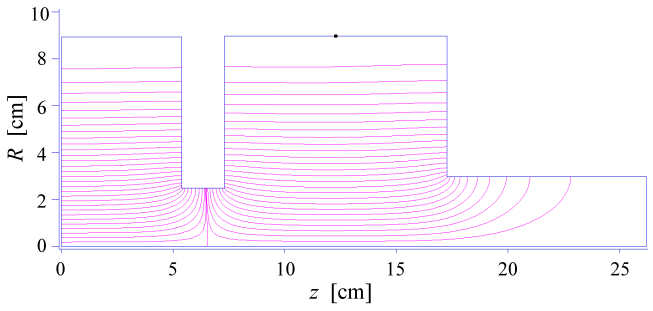

FIG. 20. Optimized straight line model of the cylindrically symmetric Photo Injector Test Facility Zeuthen (PITZ) rf gun geometry. Total length is 26.191 85 cm, and the horizontal axis is the symmetry axis. The isolines (magenta) computed by POISSON SUPERFISH show the magnetic (RH_ϕ component) field pattern for an accelerating mode.

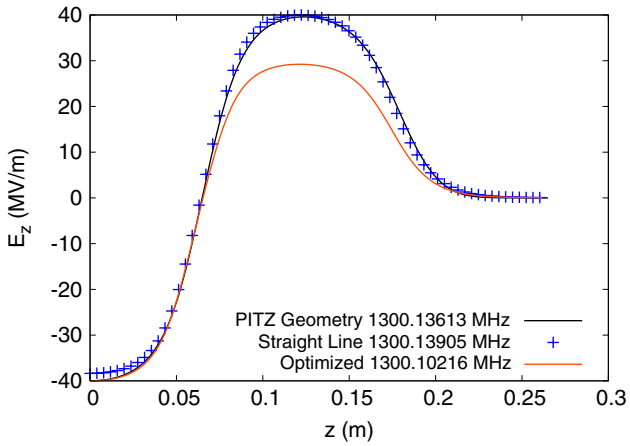

FIG. 21. PITZ on-axis balanced field profiles for the PITZ curvilinear geometry (black) and model straight line geometry (blue crosses). The optimized straight line geometry (red) is unbalanced.

is automatically invoked to produce new field profiles during optimization execution.

1. Optimization problem

Two programs, a field solver and beam dynamics simulation code, are used together to calculate the objective function values in this rf gun optimization system. APISA provides an interface to ASTRA, the beam dynamics code from DESY, and we have added one to POISSON SUPERFISH [29], a field solver for cylindrically symmetric rf cavities from Los Alamos National Laboratory. Results from ASTRA and POISSON SUPERFISH can be used as objective function values, but for the example problem presented here, only ASTRA data are used as objectives.

With the addition of the field solver to APISA, the optimization can vary user-specified aspects of the generalized geometry description shown in Fig. 22. It passes the cavity geometry description to a program that produces an on-axis accelerating or π-mode field profile. This program encapsulates the geometry description translation to POISSON SUPERFISH's format and all POISSON SUPERFISH processing required to find the on-axis π-mode. The program does not tune the cavity geometry to a desired frequency. Instead, as discussed below, constraints on frequency defined as part of the optimization problem can be used to guide the optimization to the desired frequency. In addition to the π-mode frequency, the program provides to the optimization field characteristics calculated by POISSON SUPERFISH that can be used in constraints and objectives. The generated field profile is then used in an ASTRA simulation to determine the effect on the beam dynamics.

While presently limited to straight line geometries, the generalized geometry description is versatile and can describe simple pillbox cavities as well as approximations for elliptical and reentrant cavity geometries as shown in Fig. 23. In the geometry description, each cell is described by a distinct entity or block, and a multicell cavity description is an ordered list of these blocks. This localizes each change to an individual element including changes in cell length that affect the entire cavity structure. Also, by varying, for example, the cell wall angles, a pillbox can be easily morphed into reentrant and elliptical cavity geometries, allowing optimizations that can consider all three geometries concurrently.

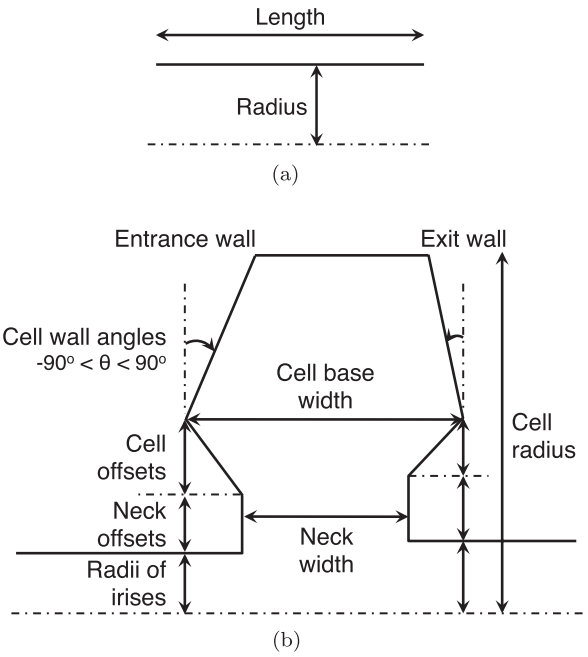

FIG. 22. Beam tube and cell parameters in general cavity description [25]: (a) beam tube or iris; (b) cell. The beam enters each element on the left.

The cavity morphing method has been used to study the PITZ rf gun injector shown in Fig. 24. This injector consists of a 1300 MHz 1.5 cell gun with a cylindrically symmetric coaxial coupler located between two solenoids [55]. The downstream solenoid is an emittance compensating solenoid, and the upstream one, known as the bucking solenoid, is used to ensure that the magnetic field at the gun cathode is zero. For production beam delivery, the bucking solenoid's field strength is set to cancel residual magnetic

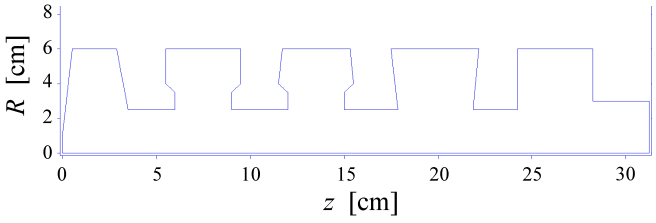

FIG. 23. Straight line approximations for elliptical (far left), reentrant (three middle), and pillbox (far right) cell geometries. Each cell radius is 6 cm, and the total length is 31.2 cm.

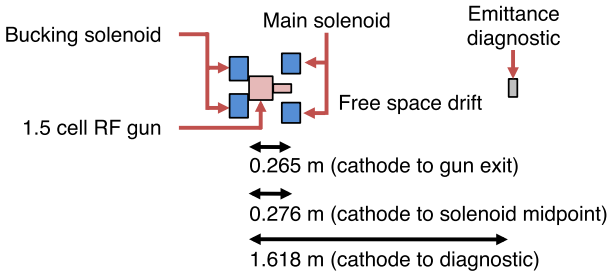

FIG. 24. Schematic of PITZ injector used in the optimization.

field from the emittance compensation solenoid at the cathode.

For this study, the injector is simplified to include only the gun and emittance compensating solenoid. Further, the coaxial coupler is omitted, and the rf gun cavity is modeled with a straight line approximation of the actual curvilinear geometry. In the curvilinear geometry, edges where two surfaces meet (e.g., outer cavity wall and cell end plate) have rounded corners whereas in the model geometry, hard corners are used as shown in Fig. 20. Figure 21 shows that the field profile for the straight line model geometry fairly well reproduces the profile from the curvilinear geometry.

The beam bunch is modeled in the optimization with a 800 pC distribution containing 2000 macroparticles. The transverse beam positions in the ASTRA macroparticle distribution [23] are uniformly distributed radially giving a 0.485 mm beam size, $\sigma_{x,y}$. The temporal profile is a 24 ps FWHM flattop pulse with 6 ps rise and fall times. The momentum distribution is isotropic with a 0.55 eV average kinetic energy corresponding to the net energy of electrons photoemitted from the gun's Cs_2Te cathode.

The optimization varies dimensions of the cavity (radii and lengths of the two cells and the intervening beam tube), the rf phase, and solenoid field strength while keeping the peak electric field fixed at 40 MV/m to find an optimal gun design and operating parameters that minimize $\varepsilon_{n,x}$, $\varepsilon_{n,y}$, $\varepsilon_{n,z}$, and the transverse beam sizes, σ_x and σ_y. Because the beam is cylindrically symmetric, the transverse dimensions are the same, so the optimization is essentially minimizing three unique quantities, σ_x, $\varepsilon_{n,x}$, and $\varepsilon_{n,z}$. The total number of independent variables, listed in Table III, is eight. The results for this 8D, three objective problem with bounded inputs and additional search space constraints, discussed next, are presented for the case of ten generations with 96 individuals per generation.

TABLE III. Injector optimization independent variables and ranges

Variable	Unit	Lower bound	Upper bound
rf phase (relative to crest)	degrees	−10	15
Main solenoid strength	Tesla	−0.180	−0.1
Iris radius	cm	2.4529	2.5029
Iris length	cm	1.9073	2.0073
Cell 1 radius	cm	8.9249	8.9449
Cell 1 length	cm	5.3763	5.4763
Cell 2 radius	cm	8.9586	8.9786
Cell 2 length	cm	9.8864	9.9864

2. Using constraints to guide search

For a multicell cavity, the resonance frequency and the field profile shape are very sensitive to small changes in the cell radii [25]. The resonance frequency for the π-mode alone is not sufficient to determine if a cavity geometry produces a reasonable field profile because for very small changes in cavity dimensions, while the frequency may remain the same, the relative field amplitudes in the field profile can change significantly. A figure of merit, such as field flatness, related to the field profile characteristics is needed to differentiate between these cases [25]. Field flatness is a gross characterization of the relative differences in peak field amplitude across a cavity. The following definition for field flatness, p_{flatness}, is used [58]:

$$p_{\text{flatness}} = 100\frac{|E_{\text{peak}}|_{\max} - |E_{\text{peak}}|_{\min}}{\frac{1}{n_{\text{cells}}}\left(\sum_{i=1}^{n_{\text{cells}}} |E_{\text{peak}}|_i\right)}, \tag{3}$$

where $|E_{\text{peak}}|_{\max}$ and $|E_{\text{peak}}|_{\min}$ are maximum and minimum peak electric field amplitudes across the cavity, $|E_{\text{peak}}|_i$ is the peak electric field amplitude in the ith cell, and n_{cells} is the number of cells, independent of their relative lengths. For a 1.5 cell cavity, n_{cells} is two, and its on-axis π-mode field profile has two peaks. Notice under this definition that a balanced field profile has a field flatness of 0%. The field flatness definition can be improved with the addition of a sign to indicate the relative order along the beam line (z) of the maximum and minimum peak field amplitudes in the cavity. This signed field flatness, p_{signed}, is defined as

$$p_{\text{signed}} = \begin{cases} -p_{\text{flatness}} & z_{|E_{\text{peak}}|_{\max}} < z_{|E_{\text{peak}}|_{\min}} \\ p_{\text{flatness}} & \text{otherwise.} \end{cases} \tag{4}$$

Using p_{signed} and APISA's strict inequality constraints, four constraints steer the optimization toward cavities with the desired frequency (1300 ± 0.5 MHz) and reasonable signed field flatness ($-101\% < p_{\text{signed}} < 101\%$). Note that for $n_{\text{cells}} = 2$, 100% field flatness means $|E_{\text{peak}}|_{\max} = 3|E_{\text{peak}}|_{\min}$ [25]. Thus, the signed field flatness constraints are not overly restrictive and allow the optimization to consider a variety of relative peak field amplitude configurations.

The progression of the average signed field flatness and frequency with generation is shown in Figs. 25 and 26. In the first generation, the individuals in the population (offspring) are randomly generated from the bounds

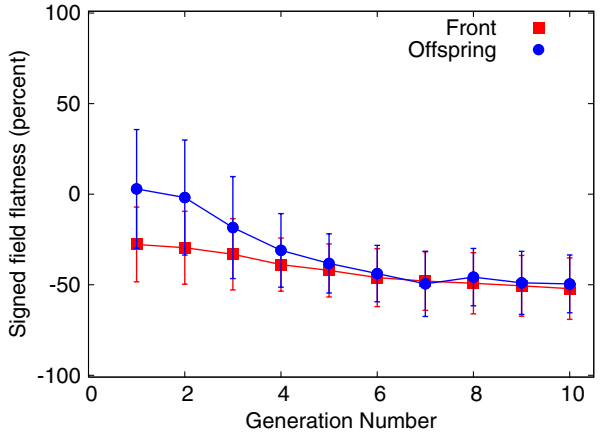

FIG. 25. Average signed field flatness for the Pareto-optimal front and offspring for each generation.

information for the independent variables. As expected, this leads to a large spread in values. For the first generation in a minimization problem, the individuals in the front are those members of the initial population that have the smallest objective values and meet the constraint limits. In subsequent generations, the front contains the individuals across the present and all past generations that have the best objective values while meeting the constraints. After the first generation, the offspring, under SPEA2 [20], are produced from the archive which contains the front and better individuals from the population. In each generation, the front has a smaller spread than the offspring, and with each generation the spread in the population decreases and eventually matches the front. Figure 26 shows that in three generations the offspring are within the 1 MHz band of acceptable frequencies, and in four generations the spread of the offspring matches the front. These figures demonstrate that the constraints are effective in guiding the optimization toward cavity geometries with the prescribed characteristics.

TABLE IV. Optimization initial and optimized straight line geometry dimensions

Geometry dimension (cm)	Initial	Optimized
Cell 1 radius	8.9349	8.933 32
Cell 1 length	5.4513	5.381 51
Iris radius	2.4779	2.482 67
Iris length	1.9823	1.912 88
Cell 2 radius	8.9686	8.972 12
Cell 2 length	9.9114	9.977 16
Exit tube radius	2.9734	2.9734
Exit tube length	8.9203	8.9203
Total length	26.2653	26.191 85

3. Results

The initial geometry dimensions are provided in Table IV. Without geometry optimization, the straight line PITZ model geometry produces the emittance product $(\varepsilon_{n,x}{}^2\varepsilon_{n,z})$ 167.88 π^3 mm^3 mrad^2 keV for the 800 pC bunch charge where $\varepsilon_{n,x} = 2.2020\ \pi$ mm mrad, $\varepsilon_{n,z} = 34.6229\ \pi$ mm keV, and $\sigma_x = 0.137\,54$ mm when the rf phase is $-2°$ off-crest and the main solenoid strength is $-0.168\,87$ T. After running ten generations with 96 individuals per generation, the optimization achieves $\varepsilon_{n,x}{}^2\varepsilon_{n,z} = 146.70\ \pi^3$ mm^3 mrad^2 keV with $\varepsilon_{n,x} = 2.1467\ \pi$ mm mrad and $\varepsilon_{n,z} = 31.834\ \pi$ mm keV. The corresponding beam size is $\sigma_x = 0.166\,49$ mm, slightly larger. The rf phase and main solenoid strength settings are comparable at $-2.25°$ off-crest and -0.1514 T, respectively. This solution represents a 13% improvement in brightness over the initial model geometry. The similarity in operational settings, rf phase, and solenoid setting, supports the conclusion that the changes in geometry are the driving force behind the improved brightness. Figure 27 shows that the objective values exhibit the same behavior as the constraints. The spread in longitudinal emittance is

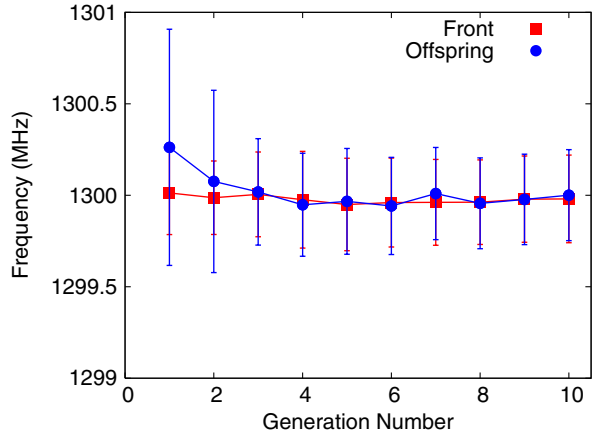

FIG. 26. Average frequency for the Pareto-optimal front and offspring for each generation.

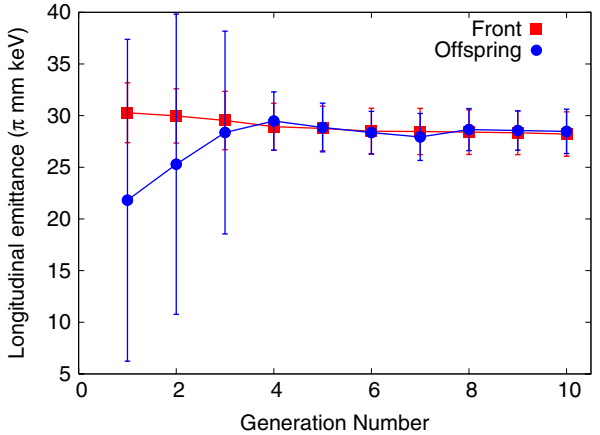

FIG. 27. Average longitudinal emittance for the Pareto-optimal front and offspring for each generation.

large initially and decreases quickly. In the fifth generation and onward, it matches the front characteristics.

Table IV and Figs. 20 and 21 provide the dimensions for the optimized geometry and its field profile. One distinctive feature of the field profile for the optimized geometry is that it is not balanced. In fact, its signed field flatness is -31.14%. Recall that from the definition for a two cell cavity negative flatness means that $|E_{\text{peak}}|_{\text{max}}$ is in the gun cell, and this is consistent with having the peak electric field on the cathode in the gun cell for smaller emittance growth [59]. However, it is contrary to the cavity design goal of 0% field flatness. Although the constraints loosely limited the range of the signed flatness to $\pm 101\%$ (notably including positive flatness), the individuals with the smallest emittances and beam size, those that constitute the front for a generation, have negative field flatness values in every generation including the first. More importantly, every front in Fig. 25 consists only of gun cavity geometries with negative field flatness values. These are all indications that better brightness sources are possible if the peak field is in the gun cell, and the field profile is unbalanced.

V. DISCUSSION AND CONCLUSION

Using GAs, we have found solutions to accelerator physics design problems that are too computationally prohibitive to be solved using standard optimization techniques. For example, the landscape of the luminosity and dynamic aperture as a function of the betatron tunes in III B is rather complicated with many local maxima due to numerous beam resonances of various orders with different relative importance. This makes a robust, global optimization with any technique other than the GA, e.g., conjugate gradient, steepest descent, etc., extremely difficult if not impossible. Each solution presented is an advancement for the field and can serve as a model for similar design problems in other machines since the GA method and our script-based tool are very general and are not tied to a specific machine type or layout. We have shown that GAs can be used to solve single-objective problems. Our multiobjective examples further outline the challenges and intricacies of multidimensional, multiobjective optimization and demonstrate the suitability of GAs to solve them. This section outlines challenges encountered in using GAs and solutions to manage them, identifies advanced problems where GAs can be used to make a difference, and provides recommendations for future development for these powerful nature-based optimization methods.

A. Challenges

GAs can manage the fine interplay between global information and local detail. They also allow for searches of the whole parameter space without any prior knowledge of favorable regions. These strengths make GAs powerful tools to solve difficult multidimensional nonlinear problems. We observe a recurring theme throughout the optimizations using GAs: the more constrained the search space—either by invoking physical reasoning, by implementing the shrinking of the search space within the algorithm itself (III C 2), or by using additional inequality constraints (IV B 2)—the faster the convergence to a near-optimal solution. It is important to properly constrain the system to succinctly and accurately restrict the search space with as much *a priori* knowledge as possible to reduce the amount of time GAs spend sampling regions of the search space that are not relevant.

GA execution time depends on two operational factors, and both have to be considered to minimize the overall optimization execution time. The first is the time for one generation to complete, and the other is the number of generations to perform. The product of these two numbers determines how long it will take the optimization to run, and both of these depend on the time to evaluate the problem model. Therefore, it is important to ensure that the model execution time is minimized and that parameter changes have deterministic effects on the problem model. The former has a clear impact on the time for a generation to complete, and the latter affects the number of generations needed. Two examples from the rf gun injector optimization application (IV B) are provided to demonstrate the effects and how they can be mitigated.

In the initial design of the rf gun-based injector optimization system, the program that encapsulates the field solver execution also tuned the cavity geometry to a given frequency. This proved to be problematic because it made the time to complete one generation difficult to predict and did not reliably produce good candidate cavity geometries. Two tuning approaches were used. The first naive method assumed independent, linear relationships between the cavity frequency and each designated cavity tuning parameter [25]. The second method used standard nonlinear optimization methods on a single-objective function, the weighted sum of the errors in frequency and signed field flatness. This failed to converge in many instances even after many iterations. Using constraints on frequency and signed field flatness removed the variability in the per-generation execution time at the expense of using additional initial generations to locate reasonable regions in the cavity parameter space as shown in Figs. 25 and 26. With the constraints approach, a generation completes in a reasonable amount of time and produces usable cavity designs.

For the optimization to succeed, it is important to ensure that the values for independent variables, objectives, and constraints have common references between individuals. This is important because the optimization draws conclusions about the relative goodness of various independent variable settings by comparing the objective values from one individual with those from the present generation and any archived individuals. If there are no common references, the changes that the optimization generates in

independent variable values and observes among individuals become arbitrary, and the model behavior is artificially randomized. This leads the optimization to perform more generations to converge or fail to converge irrespective of the number of generations performed. A simple example is using rf phase as an independent variable. If the phase reference is relative to the cavity crest phase, when the cavity generates maximum energy gain in the beam, the phase value of one individual compared to another has a physical meaning and a predictive behavior in the model. If the phase reference is internal to the simulation model, then the phase effect of one individual is random compared to the effect in another. It is important to identify such instances in an optimization problem and ameliorate them if possible.

B. Further applications of GAs

GAs are excellent tools for design optimization as the variety of challenging applications presented here exemplify. Their deceptively simplistic algorithms make them flexible and easily adaptable to machine and component design. We propose that GAs can be used to design and solve problems for next generation machines such as a muon collider and to continue to automate accelerator component design like a multifaceted cavity design process.

1. Muon colliders

There is an increasing interest in a high-luminosity high-energy muon collider to enable a new generation of fundamental particle physics exploration [60,61]. Such a machine would have the energy-frontier capability normally found in hadron colliders combined with the precision of electron-positron colliders. Designing a muon collider is an extremely challenging task. Muons are generated stochastically with a large initial 6D phase space. Because of their short lifetime, muons have to be quickly captured, cooled, accelerated, and injected into a collider ring where they must interact with an incident muon beam providing a reasonable luminosity. GA optimization is integral to essentially every aspect of a muon collider design.

One scenario for the final stage of muon beam cooling is parametric-resonance ionization cooling [62,63]. It is accomplished by inducing a half-integer parametric resonance in a muon cooling channel. The beam is then naturally focused with a period of the channel's free oscillations. The channel is designed with correlated values of the horizontal and vertical betatron periods so that focusing occurs in both planes simultaneously. Absorber plates for ionization cooling followed by energy-restoring rf cavities are placed at the beam focal points. At the absorbers, ionization cooling limits the beam angular spread while the parametric resonance causes a strong reduction of the beam spot size. The most challenging aspect of parametric-resonance ionization cooling is compensation of beam aberrations from one absorber location to another. Both chromatic and spherical aberrations must be compensated to a degree where they are small compared to the beam size at the absorber. GAs are well suited for minimizing the aberrations in this multiparameter space.

Another potential application of GAs is in designing muon recirculating linear accelerators with multipass arcs [64,65]. A return-arc optics design has been demonstrated, in which linear combined functions magnets with variable dipole and quadrupole field components are used to transport two consecutive passes with very different energies through the same string of magnets. The design requires that the arc has periodic solutions for the orbit and Twiss functions at both energies and that the periodic orbit offset, dispersion, and their slopes are all zero at the beginning and at the end of the arc at both energies. Conceptually, the number of passes through the arc is not limited to two; however, finding a solution for three or more passes is more complicated due to the increased number of independent variables and optimization criteria compounding the number of nonunique solutions to resolve. GAs can overcome or greatly simplify these optimization difficulties.

The design of a muon collider ring faces many of the same challenges as the collider rings of an electron-ion collider, only taken to extreme [66]. To make the most efficient use of expensive muons and reach desired luminosity levels, the muon beams have to be focused into very small spots at the IPs. This leads to an even greater problem with momentum acceptance and dynamic aperture than in an electron-ion collider. GAs can optimize these quantities.

2. Cavity design

GAs can still be used to improve and automate accelerator component design, especially rf and SRF structures. The primary cavity design goal is to optimize the cavity geometry to achieve the required operating frequency. Designs also need to be checked for adverse effects such as multipacting, the emission of secondary electrons from the cavity surface due to incident primary electrons that can lead to thermal breakdown, and higher-order modes that can be excited by the beam and create wakefields that degrade the beam. GAs can improve the cavity design process fundamentally by managing the complexity of the initial cavity geometry optimization. More broadly, GAs can streamline the design optimization and verification process managing the optimization of the cavity through a progression of cavity performance analysis tools to mitigate multipacting and higher-order modes.

C. Future directions: Additional capabilities

Three recommendations for increasing the power of GAs relate to problem definition and data analysis. The rf gun injector optimization shows how effective and powerful constraints are in guiding the optimization. Expanding the

types of permissible constraints can fine-tune the guidance provided by the constraints. Including equality and weak inequality constraints as well as relative constraints between characteristics in the model allows for more precise descriptions of the preferred characteristics of candidate solutions. In the course of a GA optimization, several individuals are considered, deemed unsuitable, and discarded. For problems where little is known of the search space landscape, it would be useful to have tools to analyze the individuals and generations, progressions of the front and constraints, and poorly performing solutions to learn more about the behavior of system. Finally, as studying the genome of living creatures is beneficial to understand migration of biological populations, tracing the family trees of individuals in the front may prove useful to understand how GAs arrive at the front by delineating the various paths the GA followed through the search space.

In addition to GAs, there are other nature-inspired algorithms for solving multidimensional nonlinear optimization problems. One such example is the particle-swarm algorithm (PSA) [67]. The basic strategy behind this algorithm is to mimic the dynamics of a swarm of living organisms in their quest for sustenance. By elegantly and simply appropriating the proven techniques of nature, the PSA succeeds in being easy to control, easy to understand, and potentially quite effective. Because of its simplicity, flexibility, and inherent parallelizability, the PSA has the potential to be a very useful tool for nonlinear parallel optimization.

Another promising avenue of research is creating a hybrid approach by combining the GA with traditional methods for improved convergence. The traditional, gradient-based methods, including adjoint-based methods, have the advantage of a relatively low complexity and fast convergence. The disadvantage of these gradient-based methods is that they converge to a local extremum, and as such the optimal solution depends on the initial state. The GA and PSA methods, on the other hand, are theoretically capable of converging to the global extremum. This motivates a hybrid optimization approach: first GA does the preliminary search until the problem has converged to the neighborhood of a single local extremum, followed by a gradient-based local search which provides a rapid convergence to the local extremum.

A very important property of the GAs is that they lend themselves naturally to massive parallelization. All individuals of the same generation are evaluated concurrently, and independently of each other. Depending on whether the objective function evaluator at the heart of the GA is parallelized or not, the simulations are either executed on the multi-CPU clusters or on computer farms consisting of many single-CPU nodes. However, in instances when the objective function can be executed on a hybrid platform consisting of both CPUs and graphical processing units, additional improvement in efficiency of the algorithm execution can be expected.

ACKNOWLEDGMENTS

The authors would like to express their appreciation to Ya. S. Derbenev, G. A. Krafft, T. Satogata, and Y. Zhang for their careful reviews of our paper. This paper is authored by Jefferson Science Associates, LLC under U.S. Department of Energy (DOE) Contract No. DE-AC05-06OR23177. Material presented here is also based upon work supported by the Department of Defense's ASSURE Program, the National Science Foundation under NSF Award No. 1062320, and DOE Small Business Technology Transfer Grant No. DE-SC0006272.

APPENDIX: MUTATION AND RECOMBINATION PROBABILITY DENSITY FUNCTIONS

The independent variable variations produced by mutation and recombination in these optimizations are governed by probability density functions (pdfs) with exact polynomial representations. Each pdf has a user-configurable parameter to tune the polynomial and the resulting distribution in independent variable variations. In this Appendix, we provide a brief description of the forms of the GA operators used, their associated pdfs, and tuning parameters.

1. Mutation

In mutation, the value x_o of an independent variable x is perturbed with a randomly generated small offset to produce x_{o_m}. The trend of offsets produced depends on the mutation operator employed, and these optimizations use polynomial mutation [19,68,69]. The underlying pdf has two forms, and they differ only by a normalization factor. The first form, provided to illustrate the effect of the tuning parameter $\eta_{mut} \geq 0$, is

$$p(\delta) = \tfrac{1}{2}(1 + \eta_{mut})(1 - |\delta|)^{\eta_{mut}} \tag{A1}$$

for $|\delta| \leq 1$. The simple expression to create x_{o_m} is

$$x_{o_m} = x_o + \delta\Delta_{max}, \tag{A2}$$

where Δ_{max}, set by the user, is the maximum permissible incremental change in any instance of x. In concert, Eqs. (A1) and (A2) assume the domain for independent variable x in the optimization is the entire space $-\infty \leq x \leq +\infty$, and therefore, mutation can produce any value for x_{o_m}.

Figure 28(a) shows Eq. (A1) for three values of η_{mut}. The exponent η_{mut} affects the distribution of the variations in a fairly straightforward manner. When $\eta_{mut} = 0$, the independent variable variations $\delta\Delta_{max}$ are distributed uniformly between $-\Delta_{max}$ and Δ_{max}. As η_{mut} is increased, the distribution becomes skewed toward smaller variations. In the limit as $\eta_{mut} \to \infty$, the pdf approaches a Dirac delta

function located at $\delta = 0$, $\delta_D(\delta)$, and mutation is suppressed producing $x_{o_m} = x_o$.

When an independent variable is bounded, $x^L \le x \le x^U$, the domain of δ has to be restricted to ensure that Eq. (A2) obeys the bounds on x for a bounded x_o. Judiciously rescaling Eq. (A1) with a factor that depends on x_o, x^L, and x^U adjusts the domain of δ while preserving the pdf's symmetry about $\delta = 0$. Δ_{max} is redefined in terms of the bounds to be

$$\Delta_{max} = x^U - x^L$$

and is used to define the minimum relative distance between x_o and the edge of the search space,

$$\Delta = \frac{\min[x_o - x^L, x^U - x_o]}{\Delta_{max}}.$$

Δ forms the basis of the pdf scaling factor and determines the domain of δ ($|\delta| \le \Delta$ where $0 \le \Delta \le \frac{1}{2}$). The bounded version of Eq. (A1), a function of both η_{mut} and Δ, is

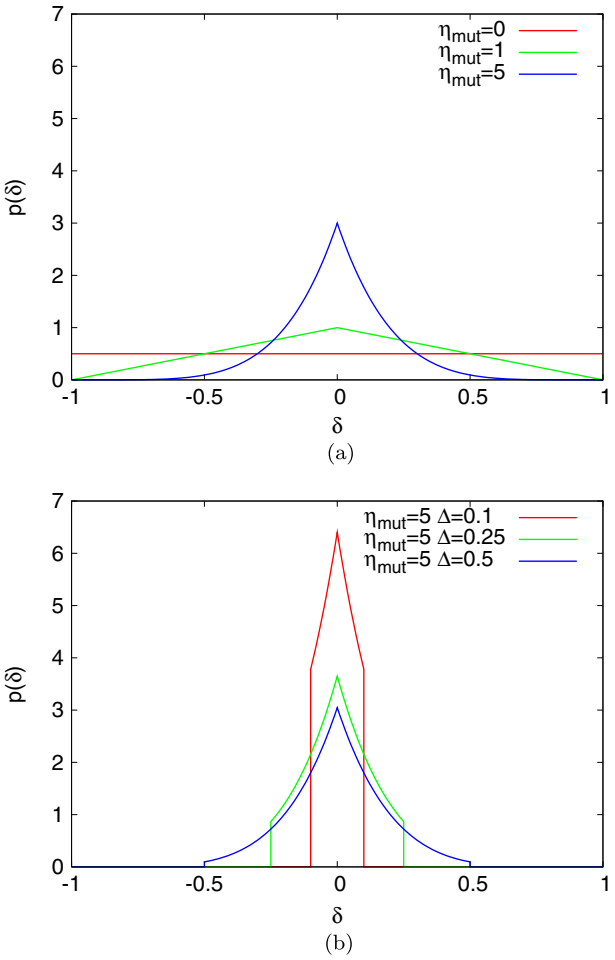

FIG. 28. Probability density functions used in polynomial mutation: (a) when the independent variable domain is unconstrained for three values of η_{mut} [Eq. (A1)]; (b) when the independent variable domain is restricted to $x^L \le x \le x^U$ [Eq. (A3)]. In (b), η_{mut} is fixed, and Δ is varied.

$$p(\delta) = \begin{cases} \delta_D(\delta) & \Delta = 0 \\ \frac{1}{2}\frac{(1+\eta_{mut})(1-|\delta|)^{\eta_{mut}}}{1-(1-\Delta)^{1+\eta_{mut}}} & 0 < \Delta \le \frac{1}{2} \end{cases} \quad \text{(A3)}$$

customized to each x_o.

Figure 28(b) shows Eq. (A3) for $\eta_{mut} = 5$ and three values of Δ. Comparing $\eta_{mut} = 5$ curves in Figs. 28(a) and 28(b), the shapes are the same, so the effect of η_{mut} is unchanged by the scaling factor. However, the extent in δ and the scaling of the curves do change with Δ demonstrating that the scaling factor is effective. When x_o is near the edge of the search space, e.g., $\Delta = 0.1$ giving $|\delta| \le 0.1$, the variation in the resulting independent variable values, $\delta\Delta_{max}$, is much smaller than when x_o is at the center of the search space ($\Delta = 0.5$ or $|\delta| \le 0.5$). This also shows that preserving the symmetry of Eq. (A1) in some ways overly restricts the variation in mutations. When x_o is one of the boundary values, $\Delta = 0$, so $x_{o_m} = x_o$ effectively turning off mutation. Arguably, in this example, x_{o_m} can take on any value in the search space, and this could be achieved if symmetry were dropped and the pdf domain shifted to $0 \le \delta \le 1$ for $x_o = x^L$ or $-1 \le \delta \le 0$ for $x_o = x^U$.

2. Recombination

Recombination, also known as *crossover*, is a pairwise operation where portions of parent genes, x_{p_1} and x_{p_2}, are exchanged through a process specific to each operator to produce offspring genes, x_{o_1} and x_{o_2}. These optimizations use simulated binary crossover (SBX) [19,68,69]. Its name alludes to the original GAs that performed operations on genes represented as binary strings, sets of bits. In the binary form, mutation flipped a bit from on to off or vice versa, and recombination exchanged sequences of bits between genes. SBX is the real valued approximation of the binary string form of recombination. To model this behavior, SBX maintains a *spread*, proportional distance, between the parents and offspring. The spread is defined as

$$\beta = \left| \frac{x_{o_1} - x_{o_2}}{x_{p_1} - x_{p_2}} \right|$$

assuming $x_{p_1} \ne x_{p_2}$. The offspring are calculated using

$$x_{o_1} = \tfrac{1}{2}\{(x_{p_1} + x_{p_2}) - \beta|x_{p_1} - x_{p_2}|\}, \quad \text{(A4)}$$

$$x_{o_2} = \tfrac{1}{2}\{(x_{p_1} + x_{p_2}) + \beta|x_{p_1} - x_{p_2}|\}. \quad \text{(A5)}$$

By design, $x_{o_1} \le x_{o_2}$, and each is equidistant from $(x_{p_1} + x_{p_2})/2$. Table V summarizes the expected results of Eqs. (A4) and (A5) for various ranges of β.

A pdf determines the distribution of β values. As with polynomial mutation, there are two forms of the pdf for bounded and unbounded independent variables. The unbounded version is

TABLE V. Correspondence for SBX between β and offspring produced for $x_{p_{low}} = \min(x_{p_1}, x_{p_2})$ and $x_{p_{high}} = \max(x_{p_1}, x_{p_2})$. In the unbounded case, $B = +\infty$, $x^L = -\infty$, and $x^U = +\infty$. For the bounded case, B is defined in Eqs. (A7) and (A8) while x^L and x^U are defined in the optimization problem statement.

β	Offspring
$\beta = 0$	$x_{o_1} = x_{o_2} = \frac{x_{p1}+x_{p2}}{2}$
$0 < \beta < 1$	$x_{p_{low}} < x_{o_1} < x_{o_2} < x_{p_{high}}$
$\beta = 1$	$x_{o_1} = x_{p_{low}}$, $x_{o_2} = x_{p_{high}}$
$1 < \beta \leq B$	$x^L \leq x_{o_1} < x_{p_{low}} < x_{p_{high}} < x_{o_2} \leq x^U$

$$p(\beta) = \frac{1}{2}(1 + \eta_{rec}) \begin{cases} \beta^{\eta_{rec}} & 0 \leq \beta \leq 1 \\ \beta^{-(2+\eta_{rec})} & 1 < \beta \end{cases} \quad \text{(A6)}$$

with tuning parameter, $\eta_{rec} \geq 0$. Unlike the polynomial mutation pdf, Eq. (A6), shown in Fig. 29(a), is not symmetric, and the domain of β does not have a finite upper bound.

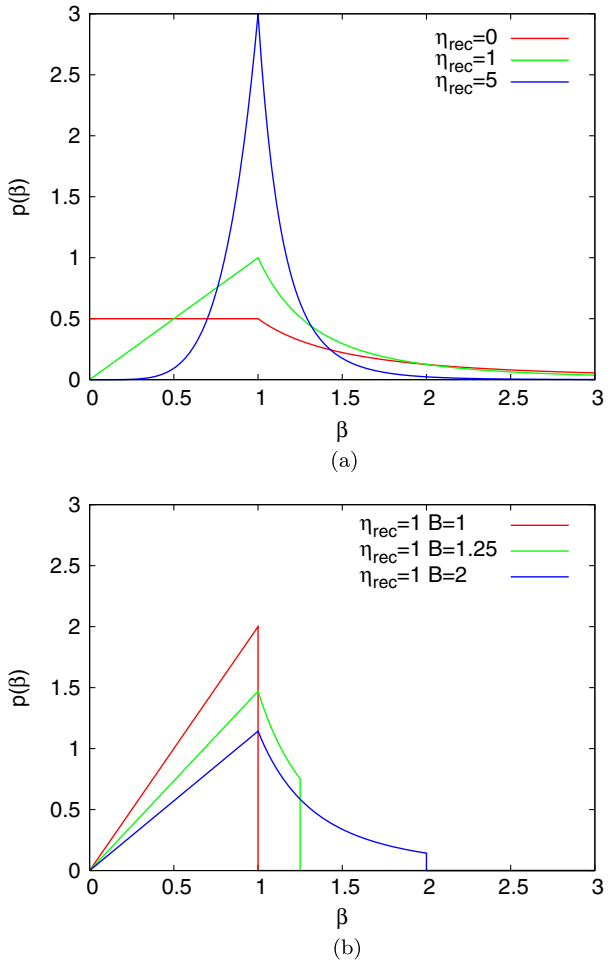

FIG. 29. Probability density functions used in simulated binary crossover recombination: (a) when the independent variable domain is unconstrained for three values of η_{rec} [Eq. (A6)]; (b) when the independent variable domain is restricted to $x^L \leq x \leq x^U$ [Eq. (A9)]. In (b), η_{rec} is fixed, and B is varied.

The effect of the tuning parameter η_{rec} parallels the influence of the polynomial mutation parameter η_{mut}. As η_{rec} is increased, the variation decreases producing offspring closer to the parent values. As $\eta_{rec} \to \infty$, the pdf approaches $\delta_D(\beta - 1)$ turning off recombination since $\beta = 1$. The pdf when $\eta_{rec} \to 0$ is broader leading to greater variation with an increasing probability of producing offspring farther away from the parents ($\beta \gg 1$). When $\eta_{rec} = 0$, the variations are uniformly distributed for $\beta \leq 1$ producing offspring between the parent values. For $\beta > 1$ when the parents are between the offspring, there is higher probability that the offspring will be close to the parents, but the probability of producing $\beta \gg 1$ with offspring far away from the parents is not zero.

When the independent variable is bounded, $x^L \leq x \leq x^U$, the pdf has to be similarly rescaled to guarantee that Eqs. (A4) and (A5) produce offspring within the independent variable bounds provided the parents obey the bounds. For $\beta \leq 1$, Eqs. (A4) and (A5) always produce offspring within the bounds, but for $\beta > 1$, there is an upper limit that depends on the relative distance between the parents and the independent variable bounds. The upper bound, B, is defined as

$$\Delta_\beta = \min[x_{p_1} - x^L, x_{p_2} - x^L, x^U - x_{p_1}, x^U - x_{p_2}] \quad \text{(A7)}$$

$$B = 1 + \frac{2\Delta_\beta}{|x_{p_2} - x_{p_1}|}. \quad \text{(A8)}$$

With a B-dependent factor Eq. (A6) is rescaled to give the bounded pdf

$$p(\beta) = \frac{1 + \eta_{rec}}{2 - B^{-(1+\eta_{rec})}} \begin{cases} \beta^{\eta_{rec}} & 0 \leq \beta \leq 1 \\ \beta^{-(2+\eta_{rec})} & 1 < \beta \leq B \end{cases} \quad \text{(A9)}$$

shown in Fig. 29(b).

B depends on the ratio of two distances: the minimum from the parents to the edges of the search space and the full distance between the parents. The ratio approaches zero as one of the parents nears an edge of the search space. This corresponds in Fig. 29(b) to the $B = 1$ curve where a parent is exactly on a search space edge, and the solutions to Eqs. (A4) and (A5) are viable only for $\beta \leq 1$. There is more than one path to producing $B \gg 1$, but the most significant one is when the parents are very close to the center of the search space, $(x^U + x^L)/2$. An example of $B \gg 1$ is the $B = 2$ curve in Fig. 29(b) allowing for greater variation in the offspring since $0 \leq \beta \leq 2$.

[1] R. Hajima, N. Takeda, H. Ohashi, and M. Akiyama, Nucl. Instrum. Methods Phys. Res., Sect. A **318**, 822 (1992).
[2] I. V. Bazarov and C. K. Sinclair, Phys. Rev. ST Accel. Beams **8**, 034202 (2005).
[3] D. Schirmer, M. v. Hartrott, S. Khan, D. Krämer, and E. Weihreter, in *Proceedings of the Particle Accelerator*

Conference, Dallas, TX, 1995 (IEEE, New York, 1995), pp. 1879–1881.

[4] D. Schirmer, P. Hartmann, T. Büning, and D. Müller, in *Proceedings of the 10th European Particle Accelerator Conference, Edinburgh, Scotland, 2006* (EPS-AG, Edinburgh, Scotland, 2006), pp. 1948–1950.

[5] A. Bacci, V. Petrillo, A. R. Rossi, and L. Serafini, in *Proceedings of the 2007 Particle Accelerator Conference, Albuquerque, New Mexico* (IEEE, New York, 2007), pp. 3295–3297.

[6] F. E. Hannon and C. Hernandez-Garcia, in Proceedings of the 10th European Particle Accelerator Conference, Edinburgh, Scotland, 2006 (Ref. [4]), pp. 3550–3552.

[7] C.-X. Wang, in *Proceedings of the 23rd Particle Accelerator Conference, Vancouver, Canada, 2009* (IEEE, Piscataway, NJ, 2009), pp. 467–469.

[8] X. Dong, C. X. Wang, A. Zholents, K.-J. Kim, and N. Sereno, in *Proceedings of the 2011 Particle Accelerator Conference, NY, USA* (IEEE, New York, 2011), pp. 2005–2007.

[9] I. V. Bazarov and H. S. Padamsee, in *Proceedings of the 21st Particle Accelerator Conference, Knoxville, 2005* (IEEE, Piscataway, NJ, 2005), pp. 1736–1738.

[10] L. Emery, in Proceedings of the 21st Particle Accelerator Conference, Knoxville, 2005 (Ref. [9]), pp. 2962–2964.

[11] Y. Wang, M. Borland, and V. Sajaev, in Proceedings of the 2011 Particle Accelerator Conference, NY, USA (Ref. [8]), pp. 787–789.

[12] S. Ramberger and S. Russenschuck, in *Proceedings of the 6th European Particle Accelerator Conference, Stockholm, 1998* (IOP, London, 1998), pp. 2014–2016.

[13] S. Russenschuck, in *Proceedings of the 6th European Particle Accelerator Conference, Stockholm, 1998* (Ref. [12]), pp. 2017–2019.

[14] Y. Xiaowei and F. Mingwu, J. Appl. Sci. (Yingyong Kexue Xuebao) **17**, 8 (1999).

[15] A. Chincarini, P. Fabbricatore, G. Gemme, R. Musenich, R. Parodi, and B. Zhang, IEEE Trans. Magn. **31**, 1566 (1995).

[16] S. Tantawi, in *Proceedings of the 48th ICFA Advanced Beam Dynamics Workshop on Future Light Sources* (SLAC National Accelerator Laboratory, Menlo Park, CA, 2010).

[17] S. Tantawi, in *Proceedings of 2011 Free Electron Conference* (EPS-AG, Shanghai, China, 2011).

[18] N. Yi, T. Dechun, and L. Yuzheng, Nucl. Instrum. Methods Phys. Res., Sect. A **462**, 356 (2001).

[19] K. Deb, *Multi-objective Optimization Using Evolutionary Algorithms*, Wiley-Interscience Series in Systems and Optimization (John Wiley and Sons, Chichester, 2001).

[20] E. Zitzler, M. Laumanns, and L. Thiele, in *Evolutionary Methods for Design, Optimisation and Control with Application to Industrial Problems (EUROGEN 2001)*, edited by K. Giannakoglou, D. Tsahalis, J. Periaux, K. Papailiou, and T. Fogarty (International Center for Numerical Methods in Engineering (CIMNE), Barcelona, Spain, 2002), pp. 95–100.

[21] S. Bleuler, M. Laumanns, L. Thiele, and E. Zitzler, Technical Report No. 154, Institut fur Technische Informatik und Kommunikationsnetze, ETH Zurich, 2002.

[22] S. Bleuler, M. Laumanns, L. Thiele, and E. Zitzler, in *Evolutionary Multi-Criterion Optimization (EMO 2003)*, Lecture Notes in Computer Science, edited by C. Fonseca, P. J. Fleming, E. Zitzler, K. Deb, and L. Thiele (Springer, New York, 2003), pp. 494–508.

[23] K. Flottmann, ASTRA: A Space Charge Tracking Algorithm [http://www.desy.de/~mpyflo].

[24] A. Hofler, P. Evtushenko, and F. Marhauser, in *Proceedings of the 2009 International Computational Accelerator Physics Conference*, edited by J. Chew (Lawrence Berkeley National Laboratory and SLAC National Accelerator Laboratory, San Francisco, CA, 2009), pp. 296–299.

[25] A. Hofler and P. Evtushenko, in Proceedings of the 2011 Particle Accelerator Conference, NY, USA (Ref. [8]), pp. 805–807.

[26] A. Hofler, Ph.D. thesis, Old Dominion University, Norfolk, VA, 2012.

[27] J. Qiang, M. A. Furman, and R. D. Ryne, Phys. Rev. ST Accel. Beams **5**, 104402 (2002).

[28] M. Borland, Technical Report No. APS LS-287, Argonne National Laboratory, 2000.

[29] J. Billen and L. M. Young, Technical Report No. LA-UR-96-1834, Los Alamos National Laboratory, 2005.

[30] Jefferson Lab LQCD Homepage [http://lqcd.jlab.org/lqcd].

[31] Jefferson Lab Scientific Computing [https://wiki.jlab.org/cc/external/wiki/index.php/Scientific_Computing_Systems].

[32] The frontiers of nuclear science: A long range plan, U.S. Department of Energy and U.S. National Science Foundation (2007).

[33] A. Afanasev, A. Bogacz, A. Bruell, L. Cardman, Y. Chao, S. Chattopadhyay, E. Chudakov, P. Degtiarenko, J. Delayen, Y. Derbenev *et al.*, Zeroth-order design report for the electron-light ion collider at CEBAF (2007) [http://web.mit.edu/eicc/DOCUMENTS/ELIC-ZDR-20070118.pdf].

[34] S. Abeyratne, A. Accardi, S. Ahmed, D. Barber, J. Bisognano, A. Bogacz, A. Castilla, P. Chevtsov, S. Corneliussen, W. Deconinck *et al.*, arXiv:1209.0757.

[35] Y. Zhang and J. Qiang, in Proceedings of the 23rd Particle Accelerator Conference, Vancouver, Canada, 2009 (Ref. [7]), pp. 2653–2655.

[36] B. Terzić and Y. Zhang, in *Proceedings of the IPAC'10 Conference, Kyoto, Japan* (ICR, Kyoto, 2010), pp. 1910–1912.

[37] B. Terzić, Y. Zhang, and J. Qiang, ICFA Beam Dynamics Newsletter, **52**, 144 (2010) [http://icfa-usa.jlab.org/archive/newsletter/icfa_bd_nl_52.pdf].

[38] Y. Derbenev, in *The Low Emittance Muon Collider Workshop* (Fermi National Accelerator Laboratory, Batavia, IL, 2007).

[39] Y. Derbenev, S. Bogacz, P. Chevtsov, A. Afanasev, C. Ankenbrandt, V. Ivanov, R. P. Johnson, and G. Wang, in Proceedings of the 23rd Particle Accelerator Conference, Vancouver, Canada, 2009 (Ref. [7]), pp. 2649–2651.

[40] V. Morozov, Y. Derbenev, F. Lin, and R. Johnson, arXiv:1208.3405v1.

[41] V. Morozov and Y. Derbenev, in *Proceedings of the 2011 International Particle Accelerator Conference*

(IPAC'11/EPS-AG, Kursaal, San Sebastián, Spain, 2011), pp. 3723–3725.

[42] F. Lin, Y. Derbenev, V. Morozov, Y. Zhang, and K. Beard, in *Proceedings of the 2012 International Particle Accelerator Conference* (IEEE, New Orleans, LA, 2012), pp. 1389–1391.

[43] J. Bengtsson, Technical Report No. SLS Note 9/97, Paul Scherrer Institut, 1997.

[44] A. Streun, Technical Report No. SLS-TME-TA-1999-0014, Paul Scherrer Institut, 1999.

[45] J. Laskar, in *Proceedings of the 20th Particle Accelerator Conference, Portland, OR, 2003* (IEEE, New York, 2003), pp. 378–382.

[46] I. Reichel, in Proceedings of the 2007 Particle Accelerator Conference, Albuquerque, New Mexico (Ref. [5]), pp. 2987–2989.

[47] C. Steier and W. Wan, in Proceedings of the IPAC'10 Conference, Kyoto, Japan (Ref. [36]), pp. 4746–4748.

[48] M. Borland, V. Sajaev, L. Emery, and A. Xiao, in Proceedings of the 23rd Particle Accelerator Conference, Vancouver, Canada, 2009 (Ref. [7]), pp. 3850–3852.

[49] L. Wang, X. Huang, Y. Nosochkov, J. A. Safranek, and M. Borland, in Proceedings of the 2012 International Particle Accelerator Conference (Ref. [42]), pp. 1380–1382.

[50] V. Lebedev, V. Nagaslaev, A. Valishev, and V. Sajaev, Nucl. Instrum. Methods Phys. Res., Sect. A **558**, 299 (2006).

[51] Y. Chao, Technical Report No. JLAB-TN-08-040, Thomas Jefferson National Accelerator Facility, 2008.

[52] C. H. Yi, S. H. Kim, M.-H. Cho, W. Namkung, H.-S. Kang, and K.-J. Kim, in Proceedings of the 2012 International Particle Accelerator Conference (Ref. [42]), pp. 1224–1226.

[53] O. J. Luiten, in *The Physics and Applications of High Brightness Electron Beams: Proceedings of the ICFA Workshop, Chia Laguna, Sardinia, 2002*, edited by J. Rosenzweig, G. A. Travish, and L. Serafini (World Scientific, Singapore, 2003), pp. 108–126.

[54] A. Hofler, P. Evtushenko, and M. Krasilnikov, in Proceedings of the 2007 Particle Accelerator Conference, Albuquerque, New Mexico (Ref. [5]), pp. 1326–1328.

[55] K. Abrahamyan, W. Ackermann, J. Bahr, I. Bohnet, J. P. Carneiro, R. Cee, K. Flottmann, U. Gensch, H. J. Grabosch, J. H. Han *et al.*, Nucl. Instrum. Methods Phys. Res., Sect. A **528**, 360 (2004).

[56] K. Tiedtke, A. Azima, N. von Bargen, L. Bittner, S. Bonfigt, S. Düsterer, B. Faatz, U. Frühling, M. Gensch, C. Gerth *et al.*, New J. Phys. **11**, 023029 (2009).

[57] I. V. Bazarov, A. Kim, M. N. Lakshmanan, and J. M. Maxson, Phys. Rev. ST Accel. Beams **14**, 072001 (2011).

[58] S. An and H. Wang, Technical Reports No. JLAB-TN-03-043 or No. SNS-NOTE-AP119, Thomas Jefferson National Accelerator Facility, and Spallation Neutron Source, Oak Ridge National Laboratory, 2003.

[59] K. McDonald, IEEE Trans. Electron Devices **35**, 2052 (1988).

[60] C. M. Ankenbrandt, M. Atac, B. Autin, V. I. Balbekov, V. D. Barger, O. Benary, J. S. Berg, M. S. Berger, E. L. Black, A. Blondel *et al.* (Muon Collider Collaboration), Phys. Rev. ST Accel. Beams **2**, 081001 (1999).

[61] M. M. Alsharo'a, C. M. Ankenbrandt, M. Atac, B. R. Autin, V. I. Balbekov, V. D. Barger, O. Benary, J. R. J. Bennett, M. S. Berger, J. S. Berg *et al.* (Muon Collider Collaboration), Phys. Rev. ST Accel. Beams **6**, 081001 (2003).

[62] Y. Derbenev and R. Johnson, in Proceedings of the 21st Particle Accelerator Conference, Knoxville, 2005 (Ref. [9]), pp. 1374–1376.

[63] Y. S. Derbenev, V. S. Morozov, A. Afanasev, K. B. Beard, R. Johnson, B. Erdelyi, and J. A. Maloney, arXiv:1205.3476.

[64] V. Morozov, S. Bogacz, Y. Roblin, K. Beard, and D. Trbojevic, in Proceedings of the 2011 Particle Accelerator Conference, NY, USA (Ref. [8]), pp. 196–198.

[65] V. S. Morozov, S. A. Bogacz, Y. R. Roblin, and K. B. Beard, Phys. Rev. ST Accel. Beams **15**, 060101 (2012).

[66] Y. I. Alexahin, E. Gianfelice-Wendt, V. V. Kashikhin, N. V. Mokhov, A. V. Zlobin, and V. Y. Alexakhin, Phys. Rev. ST Accel. Beams **14**, 061001 (2011).

[67] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, Pattern Recogn. Lett. **28**, 459 (2007).

[68] K. Deb and S. Agrawal, in *Artificial Neural Nets and Genetic Algorithms: Proceedings of the International Conference in Portorož, Slovenia, 1999* (Springer, New York, 1999), pp. 235–243.

[69] K. Deb, Comput. Methods Appl. Mech. Eng. **186**, 311 (2000).