



AI-assisted optimization of the ECCE tracking system at the Electron Ion Collider



C. Fanelli^{34,84,27,*}, Z. Papandreou⁷⁴, K. Suresh⁷⁴, J.K. Adkins³³, Y. Akiba^{50,54}, A. Albataineh⁶⁹, M. Amarian⁴⁴, I.C. Arsene⁷³, C. Ayerbe Gayoso³⁵, J. Bae⁵⁹, X. Bai⁷⁹, M.D. Baker^{5,27}, M. Bashkanov⁸⁸, R. Bellwied⁶⁷, F. Benmokhtar¹⁵, V. Berdnikov¹³, J.C. Bernauer^{52,53,54}, F. Bock⁴⁶, W. Boeglin¹⁶, M. Borysova⁸³, E. Brash¹¹, P. Brindza²⁷, W.J. Briscoe²⁰, M. Brooks³⁰, S. Bueltmann⁴⁴, M.H.S. Bukhari²⁶, A. Bylinkin⁶⁹, R. Capobianco⁶⁵, W.-C. Chang², Y. Cheon⁵⁷, K. Chen⁸, K.-F. Chen⁴³, K.-Y. Cheng³⁷, M. Chiu⁵, T. Chujo⁷⁶, Z. Citron⁴, E. Cline^{52,53}, E. Cohen⁴¹, T. Cormier^{46,a}, Y. Corrales Morales³⁰, C. Cotton⁷⁹, J. Crafts¹³, C. Crawford⁷⁰, S. Creekmore⁴⁶, C. Cuevas²⁷, J. Cunningham⁴⁶, G. David⁵, C.T. Dean³⁰, M. Demarteau⁴⁶, S. Diehl⁶⁵, N. Doshita⁸⁵, R. Dupré²², J.M. Durham³⁰, R. Dzhygadlo¹⁹, R. Ehlers⁴⁶, L. El Fassi³⁵, A. Emmert⁷⁹, R. Ent²⁷, R. Fatemi⁷⁰, S. Fegan⁸⁸, M. Finger⁹, M. Finger Jr⁹, J. Frantz⁴⁵, M. Friedman²¹, I. Friscic^{34,27}, D. Gangadharan⁶⁷, S. Gardner¹⁸, K. Gates¹⁸, F. Geurts⁴⁹, R. Gilman⁵¹, D. Glazier¹⁸, E. Glimos^{46,78}, Y. Goto^{50,54}, N. Grau³, S.V. Greene⁸⁰, A.Q. Guo²⁴, L. Guo¹⁶, S.K. Ha⁸⁶, J. Haggerty⁵, T. Hayward⁶⁵, X. He¹⁷, O. Hen³⁴, D.W. Higinbotham²⁷, M. Hoballah²², T. Horn¹³, A. Hoggmsryan¹, P.-h.J. Hsu⁴², J. Huang⁵, G. Huber⁷⁴, A. Hutson⁶⁷, K.Y. Hwang⁸⁷, C.E. Hyde⁴⁴, M. Inaba⁶³, T. Iwata⁸⁵, H.S. Jo²⁹, K. Joo⁶⁵, N. Kalantarians⁸¹, G. Kalicy¹³, K. Kawade⁵⁸, S.J.D. Kay⁷⁴, A. Kim⁶⁵, B. Kim⁵⁹, C. Kim⁴⁸, M. Kim⁵⁰, Y. Kim⁴⁸, Y. Kim⁵⁷, E. Kistenev⁵, V. Klimenko⁶⁵, S.H. Ko⁵⁶, I. Korover³⁴, W. Korsch⁷⁰, G. Krintiras⁶⁹, S. Kuhn⁴⁴, C.-M. Kuo³⁷, T. Kutz³⁴, J. Lajoie²⁵, D. Lawrence²⁷, S. Lebedev²⁵, H. Lee⁵⁹, J.S.H. Lee⁷⁵, S.W. Lee²⁹, Y.-J. Lee³⁴, W. Li⁴⁹, W.B. Li^{52,53,84}, X. Li⁶¹, X. Li¹⁰, X. Li³⁰, X. Li³⁴, Y.T. Liang²⁴, S. Lim⁴⁸, C.-H. Lin², D.X. Lin²⁴, K. Liu³⁰, M.X. Liu³⁰, K. Livingston¹⁸, N. Liyanage⁷⁹, W.J. Llope⁸², C. Loizides⁴⁶, E. Long⁷², R.-S. Lu⁴³, Z. Lu¹⁰, W. Lynch⁸⁸, S. Mantry⁶⁶, D. Marchand²², M. Marcisovsky¹⁴, C. Markert⁷⁷, P. Markowitz¹⁶, H. Marukyan¹, P. McGaughey³⁰, M. Mihovilovic⁷¹, R.G. Milner³⁴, A. Milov⁸³, Y. Miyachi⁸⁵, A. Mkrtchyan¹, P. Monaghan¹¹, R. Montgomery¹⁸, D. Morrison⁵, A. Movsisyan¹, H. Mkrtchyan¹, A. Mkrtchyan¹, C. Munoz Camacho²², M. Murray⁶⁹, K. Nagai³⁰, J. Nagle⁶⁴, I. Nakagawa⁵⁰, C. Nattrass⁷⁸, D. Nguyen²⁷, S. Niccolai²², R. Nouicer⁵, G. Nukazuka⁵⁰, M. Nycz⁷⁹, V.A. Okorokov⁴⁰, S. Orešić⁷⁴, J.D. Osborn⁴⁶, C. O'Shaughnessy³⁰, S. Paganis⁴³, S.F. Pate³⁹, M. Patel²⁵, C. Paus³⁴, G. Penman¹⁸, M.G. Perdekamp⁶⁸, D.V. Perepelitsa⁶⁴, H. Periera da Costa³⁰, K. Peters¹⁹, W. Phelps¹¹, E. Piasetzky⁶⁰, C. Pinkenburg⁵, I. Prochazka⁹, T. Protzman³¹, M.L. Purschke⁵, J. Putschke⁸², J.R. Pybus³⁴, R. Rajput-Ghoshal²⁷, J. Rasson⁴⁶, B. Raue¹⁶, K.F. Read^{46,78}, K. Røed⁷³, R. Reed³¹, J. Reinhold¹⁶, E.L. Renner³⁰, J. Richards⁶⁵, C. Riedl⁶⁸, T. Rinn⁵, J. Roche⁴⁵, G.M. Roland³⁴, G. Ron²¹, M. Rosati²⁵, C. Royon⁶⁹, J. Ryu⁴⁸, S. Salur⁵¹, N. Santiesteban³⁴, R. Santos⁶⁵, M. Sarsour¹⁷, J. Schambach⁴⁶, A. Schmidt²⁰, N. Schmidt⁴⁶, C. Schwarz¹⁹, J. Schwiening¹⁹, R. Seidl^{50,54}, A. Sickles⁶⁸, P. Simmerling⁶⁵, S. Sirca⁷¹, D. Sharma¹⁷, Z. Shi³⁰, T.-A. Shibata³⁸, C.-W. Shih³⁷, S. Shimizu⁵⁰, U. Shrestha⁶⁵, K. Slifer⁷²,

* Corresponding author at: The College of William and Mary, Williamsburg, VA, USA.

E-mail address: cfanelli@wm.edu (C. Fanelli).

^a Deceased.

K. Smith³⁰, D. Sokhan^{18,23}, R. Soltz³², W. Sondheim³⁰, J. Song¹⁰, J. Song⁴⁸, I.I. Strakovsky²⁰, P. Steinberg⁵, P. Stepanov¹³, J. Stevens⁸⁴, J. Strube⁴⁷, P. Sun¹⁰, X. Sun⁸, V. Tadevosyan¹, W.-C. Tang³⁷, S. Tapia Araya²⁵, S. Tarafdar⁸⁰, L. Teodorescu⁶, D. Thomas⁷⁷, A. Timmins⁶⁷, L. Tomasek¹⁴, N. Trotta⁶⁵, R. Trotta¹³, T.S. Tveter⁷³, E. Umaka²⁵, A. Usman⁷⁴, H.W. van Hecke³⁰, C. Van Hulse²², J. Velkovska⁸⁰, E. Voutier²², P.K. Wang²², Q. Wang⁶⁹, Y. Wang⁸, Y. Wang⁶², D.P. Watts⁸⁸, N. Wickramaarachchi¹³, L. Weinstein⁴⁴, M. Williams³⁴, C.-P. Wong³⁰, L. Wood⁴⁷, M.H. Wood⁷, C. Woody⁵, B. Wyslouch³⁴, Z. Xiao⁶², Y. Yamazaki²⁸, Y. Yang³⁶, Z. Ye⁶², H.D. Yoo⁸⁷, M. Yurov³⁰, N. Zachariou⁸⁸, W.A. Zajc¹², W. Zha⁶¹, J.-L. Zhang⁵⁵, J.-X. Zhang⁷⁹, Y. Zhang⁶², Y.-X. Zhao²⁴, X. Zheng⁷⁹, P. Zhuang⁶²

¹ A. Alikhanyan National Laboratory, Yerevan, Armenia² Institute of Physics, Academia Sinica, Taipei, Taiwan³ Augustana University, Sioux Falls, SD, USA⁴ Ben-Gurion University of the Negev, Beer-Sheva, Israel⁵ Brookhaven National Laboratory, Upton, NY, USA⁶ Brunel University London, Uxbridge, UK⁷ Canisius College, Buffalo, NY, USA⁸ Central China Normal University, Wuhan, China⁹ Charles University, Prague, Czech Republic¹⁰ China Institute of Atomic Energy, Fangshan, Beijing, China¹¹ Christopher Newport University, Newport News, VA, USA¹² Columbia University, New York, NY, USA¹³ Catholic University of America, Washington DC, USA¹⁴ Czech Technical University, Prague, Czech Republic¹⁵ Duquesne University, Pittsburgh, PA, USA¹⁶ Florida International University, Miami, FL, USA¹⁷ Georgia State University, Atlanta, GA, USA¹⁸ University of Glasgow, Glasgow, UK¹⁹ GSI Helmholtzzentrum fuer Schwerionenforschung GmbH, Darmstadt, Germany²⁰ The George Washington University, Washington DC, USA²¹ Hebrew University, Jerusalem, Israel²² Universite Paris-Saclay, CNRS/IN2P3, IJCLab, Orsay, France²³ IRFU, CEA, Universite Paris-Saclay, Gif-sur-Yvette, France²⁴ Chinese Academy of Sciences, Lanzhou, China²⁵ Iowa State University, Iowa City, IA, USA²⁶ Jazan University, Jazan, Saudi Arabia²⁷ Thomas Jefferson National Accelerator Facility, Newport News, VA, USA²⁸ Kobe University, Kobe, Japan²⁹ Kyungpook National University, Daegu, Republic of Korea³⁰ Los Alamos National Laboratory, Los Alamos, NM, USA³¹ Lehigh University, Bethlehem, PA, USA³² Lawrence Livermore National Laboratory, Livermore, CA, USA³³ Morehead State University, Morehead, KY, USA³⁴ Massachusetts Institute of Technology, Cambridge, MA, USA³⁵ Mississippi State University, Mississippi State, MS, USA³⁶ National Cheng Kung University, Tainan, Taiwan³⁷ National Central University, Chungli, Taiwan³⁸ Nihon University, Tokyo, Japan³⁹ New Mexico State University, Las Cruces, NM, USA⁴⁰ National Research Nuclear University MEPhI, Moscow, Russian Federation⁴¹ Nuclear Research Center - Negev, Beer-Sheva, Israel⁴² National Tsing Hua University, Hsinchu, Taiwan⁴³ National Taiwan University, Taipei, Taiwan⁴⁴ Old Dominion University, Norfolk, VA, USA⁴⁵ Ohio University, Athens, OH, USA⁴⁶ Oak Ridge National Laboratory, Oak Ridge, TN, USA⁴⁷ Pacific Northwest National Laboratory, Richland, WA, USA⁴⁸ Pusan National University, Busan, Republic of Korea⁴⁹ Rice University, Houston, TX, USA⁵⁰ RIKEN Nishina Center, Wako, Saitama, Japan⁵¹ The State University of New Jersey, Piscataway, NJ, USA⁵² Center for Frontiers in Nuclear Science, Stony Brook, NY, USA⁵³ Stony Brook University, Stony Brook, NY, USA⁵⁴ RIKEN BNL Research Center, Upton, NY, USA⁵⁵ Shandong University, Qingdao, Shandong, China⁵⁶ Seoul National University, Seoul, Republic of Korea⁵⁷ Sejong University, Seoul, Republic of Korea⁵⁸ Shinshu University, Matsumoto, Nagano, Japan⁵⁹ Sungkyunkwan University, Suwon, Republic of Korea⁶⁰ Tel Aviv University, Tel Aviv, Israel⁶¹ University of Science and Technology of China, Hefei, China⁶² Tsinghua University, Beijing, China⁶³ Tsukuba University of Technology, Tsukuba, Ibaraki, Japan⁶⁴ University of Colorado Boulder, Boulder, CO, USA⁶⁵ University of Connecticut, Storrs, CT, USA

⁶⁶ University of North Georgia, Dahlonega GA, USA⁶⁷ University of Houston, Houston, TX, USA⁶⁸ University of Illinois, Urbana, IL, USA⁶⁹ University of Kansas, Lawrence, KS, USA⁷⁰ University of Kentucky, Lexington, KY, USA⁷¹ University of Ljubljana, Ljubljana, Slovenia⁷² University of New Hampshire, Durham, NH, USA⁷³ University of Oslo, Oslo, Norway⁷⁴ University of Regina, Regina, SK, Canada⁷⁵ University of Seoul, Seoul, Republic of Korea⁷⁶ University of Tsukuba, Tsukuba, Japan⁷⁷ University of Texas, Austin, TX, USA⁷⁸ University of Tennessee, Knoxville, TN, USA⁷⁹ University of Virginia, Charlottesville, VA, USA⁸⁰ Vanderbilt University, Nashville, TN, USA⁸¹ Virginia Union University, Richmond, VA, USA⁸² Wayne State University, Detroit, MI, USA⁸³ Weizmann Institute of Science, Rehovot, Israel⁸⁴ The College of William and Mary, Williamsburg, VA, USA⁸⁵ Yamagata University, Yamagata, Japan⁸⁶ Yarmouk University, Irbid, Jordan⁸⁷ Yonsei University, Seoul, Republic of Korea⁸⁸ University of York, York, UK

ARTICLE INFO

Keywords:

ECCE

Electron Ion Collider

Tracking

Artificial Intelligence

Evolutionary algorithms

Bayesian optimization

ABSTRACT

The Electron-Ion Collider (EIC) is a cutting-edge accelerator facility that will study the nature of the “glue” that binds the building blocks of the visible matter in the universe. The proposed experiment will be realized at Brookhaven National Laboratory in approximately 10 years from now, with detector design and R&D currently ongoing. Notably, EIC is one of the first large-scale facilities to leverage Artificial Intelligence (AI) already starting from the design and R&D phases. The EIC Comprehensive Chromodynamics Experiment (ECCE) is a consortium that proposed a detector design based on a 1.5 T solenoid. The EIC detector proposal review concluded that the ECCE design will serve as the reference design for an EIC detector. Herein we describe a comprehensive optimization of the ECCE tracker using AI. The work required a complex parametrization of the simulated detector system. Our approach dealt with an optimization problem in a multidimensional design space driven by multiple objectives that encode the detector performance, while satisfying several mechanical constraints. We describe our strategy and show results obtained for the ECCE tracking system. The AI-assisted design is agnostic to the simulation framework and can be extended to other sub-detectors or to a system of sub-detectors to further optimize the performance of the EIC detector.

1. Introduction

The Electron Ion Collider (EIC) [1] is a future cutting-edge discovery machine that will unlock the secrets of the gluonic force binding the building blocks of the visible matter in the universe. The EIC will consist of two intersecting accelerators, one producing an intense beam of electrons and the other a beam of protons or heavier atomic nuclei; it will be the only electron–nucleus collider operating in the world. The EIC Comprehensive Chromodynamics Experiment (ECCE) [2] is an international consortium assembled to develop a detector that can offer full energy coverage and an optimized far forward detection region. ECCE has investigated a detector design based on the existing BABAR 1.5T magnet; this detector will be ready for the beginning of EIC operations. More details on the ECCE detector design and what is described in the following can be found in [3].

ECCE is an integrated detector that extends for about 40 m, and includes a central detector built around the interaction point and far-forward (hadron-going direction) and far-backward (electron-going direction) regions [1]. To fulfill the physics goals of the EIC, the central detector needs to be hermetic and provide good particle identification (PID) over a large phase space. The central detector itself consists of multiple sub-detectors: a tracking system made by inner and outer tracker stations allows the reconstruction of charged particles moving in the magnetic field; a system of PID sub-detectors will cover the barrel and the electron-going and hadron-going directions; electromagnetic and hadronic calorimeters are used to detect showers and provide complete information on the particle flow which is essential for certain event topologies, e.g., those containing jets.

As outlined in [1], Artificial Intelligence (AI) can provide dedicated strategies for complex combinatorial searches and can handle multi-objective problems characterized by a multidimensional design space, allowing the identification of hidden correlations among the design parameters. ECCE included these techniques in the design workflow during the detector proposal. At first this AI-assisted design strategy was used to steer the design. After the base technology is selected using insights provided by AI, its detector parameters can be further fine-tuned using AI. During the ECCE detector proposal stage, the design of the detector underwent a continual optimization process [4].

The article is structured as follows: in Section 2 we provide an overview of design optimization and describe the AI-assisted strategy; in Section 3 we introduce the ECCE tracker and describe the software stack utilized in this work to which AI is coupled for the optimization; in Section 4 we describe the implemented pipeline that results in a sequential strategy, fostering the interplay between the different working groups in a post hoc decision making process; in Section 5 we present perspectives and planned activities.

The ECCE detector at the EIC will be one of the first examples of detectors that will be realized leveraging AI during the design and R&D phases.

2. AI-assisted detector design

Detector optimization with AI is anticipated to continue in the months following the detector proposal towards CD-2 and CD-3. Optimizing the design of large-scale detectors such as ECCE – that are made of multiple sub-detector systems – is a complex problem. Each sub-detector system is characterized by a multi-dimensional design

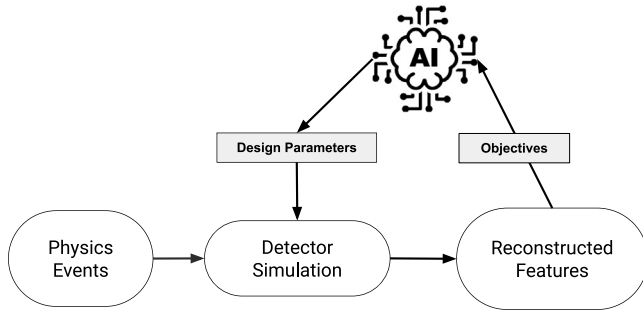


Fig. 1. Workflow of detector design assisted by AI: physics events are injected in a detector characterized by given design parameters. Reconstructed events are analyzed and figures of merit are quantified and passed to an AI-based strategy, which in turn suggests the next design point in this sequential approach; note that AI can also intervene in the simulation and reconstruction steps.

parameter space. In addition, detector simulations are typically computationally intensive, and rely on advanced simulation platforms used in our community such as GEANT4 [5] to simulate the interaction of radiation with matter. Additional computationally expensive steps are present along the data reconstruction and analysis pipeline. The software stack that is utilized in the detector design process involves three main steps: (i) generation of events, (ii) detector simulations and (iii) reconstruction and analysis.

As pointed out in [6], the above bottlenecks render the generation and exploration of multiple design points cumbersome. This in turn represents an obstacle for deep learning (DL)-based approaches that learn the mapping between the design space and the functional space [7–9], which could facilitate the identification of optimal design points. In principle fast simulations with DL can reduce the most CPU-intensive parts of the simulation and provide accurate results [10], although several design points need to be produced with GEANT4 before injection in any DL architecture. Similar considerations exist in deploying DL for reconstruction during the design optimization process.

In this context, a workflow for detector design that has gained popularity in recent years [11] is represented by the schematic in Fig. 1. It consists of a sequential AI-based strategy that collects information associated with previously generated design points, in the form of figures of merit (called *objectives* in the following) that quantify the goodness of the design and suggest promising new design points for the next iteration.

The ECCE AI Working Group achieved a continual multi-objective optimization (MOO) of the tracker design. Our approach deals with a complex optimization in multidimensional design space (describing, e.g., geometry, mechanics, optics, etc.) driven by multiple objectives that encode the detector performance while satisfying several mechanical constraints. This framework has been developed in a way that can be easily extended to other sub-detectors or to a system of sub-detectors.

The definition of a generic MOO problem can be formulated as follows:

$$\begin{aligned}
 \min \mathbf{f}_m(\mathbf{x}) \quad & m = 1, \dots, M \\
 \text{s.t.} \quad & \mathbf{g}_j(\mathbf{x}) \leq 0, \quad j = 1, \dots, J \\
 & \mathbf{h}_k(\mathbf{x}) = 0, \quad k = 1, \dots, K \\
 & x_i^L \leq x_i \leq x_i^U, \quad i = 1, \dots, N
 \end{aligned} \quad (1)$$

where one has M objective functions f_m to optimize (e.g., detector resolution, efficiency, costs), subject to J inequalities $g_j(x)$ and K equality constraints $h_k(z)$ (e.g., mechanical constraints), in a design space of N dimensions (e.g., geometry parameters that change the GEANT4 design) with lower and upper bounds on each dimension.¹ Notice that overlaps

¹ Constraints are described later in Table 3.

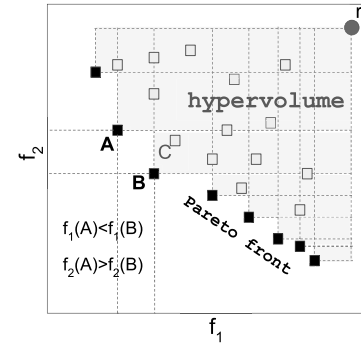


Fig. 2. Example of Pareto frontier in a two-dimensional objective space: Point C is not on the frontier and is dominated by both point A and point B. All the other points which are dominated by the Pareto frontier and that satisfy any constraints in the optimization problem are called *feasible* solutions; The hypervolume is used as a metric for convergence and is calculated with respect to a reference point r .

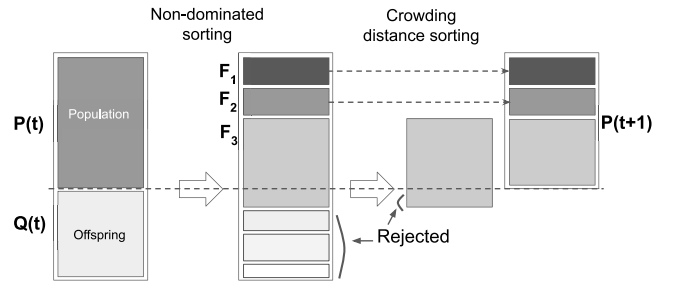


Fig. 3. The NSGA-II Workflow: At time t , an offspring $Q(t)$ is created through a genetic algorithm [15] from an N -sized population of design points $P(t)$. The two populations are combined into an augmented population which is classified into different non-dominated fronts F_i (Fig. 2), starting from the first front F_1 . To restore the initial size of the population, the augmented space of solutions is trimmed. A metric called crowding distance is used to reject solutions from the least performing front and eventually provide an updated population of size N at time $t + 1$.

in the design are checked before and during the optimization and are excluded by the constraints and ranges of the parameters. In solving these problems, one can come up with a set of *non-dominated* or *trade-off* solutions [12], popularly known as Pareto-optimal solutions (see also Fig. 2).

In this setting, we used a recently developed framework for MOO called pymoo [13] which supports evolutionary MOO algorithms such as Non-Dominated Sorting Genetic Algorithm (or NSGA-II, [14]).² The rationale behind this choice instead of, for example, principled approaches such as Bayesian Optimization [11], emanates from the ECCE needs at the time of the detector proposal, such as the capability to quickly implement and run multiple parallel optimization pipelines implementing different technology choices and the possibility of dealing with non-differentiable objectives at the exploratory stage.

The main features of NSGA-II are (i) the usage of an elitist principle that retains the best individuals, (ii) an explicit diversity preserving mechanism, and (iii) ability of determining non-dominated solutions. The latter feature is of great importance for problems where objectives are in conflict with each other: that is an improved performance in an objective results in worse performance in another objective.

The NSGA-II workflow is described in Fig. 3 and can be briefly summarized in the following steps:³

(i) *Initial population generation*: a random population P_0 of N design points is created. The population is sorted based on non-domination

² The pymoo framework also supports other MOO approaches and a full list is documented in [13].

³ The readers are directed to [13,14,16] for a comprehensive explanation of the algorithm.

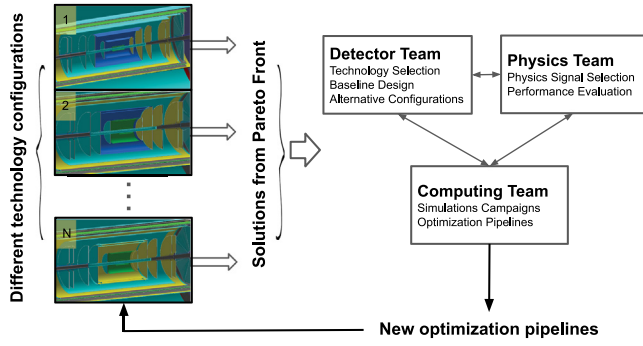


Fig. 4. Flowchart of continual optimization during proposal: AI assists the design optimization process by providing insights and capturing hidden correlations among the design parameters. This has been used during the entire detector proposal process to steer the design. At a given instant in time, N alternative candidate configurations are studied. For each, we create an optimization pipeline, which results in a Pareto front of solutions. This new information helps steer the design: some configurations are rejected, while other ones (also dubbed ‘new references’) are identified to potentially improve the design. New optimization pipelines are defined inspired by the new results and the process is iterated. During the design process, AI propelled the fundamental interplay between the ECCE Teams working on Physics, Detector and Computing.

as briefly described in Fig. 2. Each solution is assigned a fitness value equal to its non domination-level (see [16]). (ii) *Creation of offspring population:* an offspring (also called child) population is created using standard operators of genetic algorithms [15], namely binary tournament selection, crossover, and mutation operators; this creates a child population Q_0 . (iii) *Sorting the combined population:* the combined population of design points is characterized by the set $R_0 = P_0 \cup Q_0$. The same procedure is applied at any time t , so this can be generalized to $R_t = P_t \cup Q_t$. The combined population R_t is sorted according to non-domination. (iv) *Creation of new population through crowding distance:* the new parent population P_{t+1} is formed by gathering solutions from the first front and continuing to the other fronts (F_1, F_2, \dots) until the size of P_{t+1} exceeds the initial size N with the inclusion of the last acceptable front. A metric called *crowding distance* [14] is used to select the solutions of the last accepted front and trim all the other solutions, in such a way that P_{t+1} eventually consists of N solutions, i.e., it has the same size of P_t . It can be shown that the crowding distance allows to keep those points in the last front that contribute to the highest diversity [14].

For our purposes, we also tested NSGA-III which is suitable for the optimization of large number of objectives [17].⁴

During the design optimization process of the tracking system, we used full GEANT4 simulations of the entire ECCE detector. AI played a crucial role in helping choose a combination of technologies for the inner tracker and was used as input to multiple iterations of the ECCE tracker design, which led to the current tracker layout. This was the result of a continual optimization process that evolved in time: results were validated by looking at figures of merit that do not enter as objective functions in the optimization process (more details can be found in Appendix B); the decision making is left post hoc and discussed among the Computing, Detector and Physics teams. A flowchart describing this continual optimization process is shown in Fig. 4.

Ultimately this continual AI-assisted optimization led to a projective design after having extended the parametrized design to include the support structure of the inner tracker. The latter represents an ongoing R&D project that is discussed in the next sections.

3. ECCE tracking system simulation

The simulation and detector response shown in this document is based on GEANT4 [18] and was carried out using the Fun4All framework [19,20].

The optimization pipelines are based on particle gun samples of pions, where we used π^- and tested that the performance with π^+ were consistent. Performance in the electron-going direction was also checked post-hoc with particle gun samples of electrons. The improved performance is further validated with physics analyses, using the datasets generated during the ECCE simulation campaigns; in Section 4 we show in particular results based on semi-inclusive deep inelastic scattering (SIDIS) events.

The ECCE tracking detector [21], represented in Fig. 5 (left), consists of different layers in the barrel and the two end-caps, and is tightly integrated with the PID detectors:

(i) The silicon vertex/tracking detector is an ALICE ITS-3 type high precision cylindrical/disk vertex tracker [22,23] based on the new Monolithic Active Pixel Sensor (MAPS); the barrel detector consists of 5 MAPS layers (3 ITS3 Vertex ITS3 layers, 2 ITS3 Sagitta layers); the silicon hadron endcap consists of 5 MAPS disks (FST — Forward Silicon Trackers); and the silicon electron endcap (EST — Electron Silicon Trackers) has 4 MAPS disks.

(ii) A gas tracking system is based on μ Rwell technology, that is a single-stage amplification Micro Pattern Gaseous Detector (MPGD) that is a derivative of the Gas Electron Multiplier (GEM) technology. In ECCE μ Rwell layers will form three barrel tracking layers further out from the beam-pipe than the silicon layers; namely, two inner-barrel layers and a single outer-barrel μ Rwell layer. All μ Rwell detectors will have 2D strip based readout. The strip pitch for all three layers will be 400 μ m.

(iii) The tracking system is completed by AC-LGAD-based (TTL — Timing Tracking Layers [24]) time of flight (TOF) detectors providing additional hit information for track reconstruction as well. In the central region, a TOF (dubbed CTTL-Central TTL) is placed behind the high-performance DIRC (hpDIRC); in the hadron-going side a TOF (dubbed FTTL-Forward TTL) is placed before the dual RICH (dRICH) and a μ Rwell placed after the dRICH; in the electron-going direction, a μ Rwell layer is placed before the modular RICH (mRICH), which is followed by a TOF later (dubbed ETTL-Electron TTL).

An important consideration for all large-scale detectors is the provision of readout (power and signal cables) and other services (e.g., cooling). Clearly the aim is to minimize the impact of readout and services in terms of affecting the detector’s acceptance or tracking resolution, for example. This effort is ongoing R&D for the project.

In the following sections, the reader can find more details on the implementation of the optimization pipelines and utilized computing resources.

4. Analysis Workflow

The optimization of the ECCE-tracking system [3,21] has been characterized by two main phases during which the sub-detectors composing the tracker evolved into more advanced renditions.

Phase I optimization.⁵ The GEANT4 implementation of the detectors were at first simplified, e.g., detector modules were mounted on a simplified conical support structure made of aluminum. The optimization pipelines consisted of symmetric arrangement of detectors in the electron-going and hadron-going directions (5 disks on each side). The DIRC detector for PID in the barrel region was modeled with a simple geometry made by a cylinder and conical mirrors. AC-LGAD-based

⁴ For ≥ 4 objectives, NSGA-III is expected to perform better than NSGA-II.

⁵ Phase I corresponds to a timeline between June-2021 to Sept-2021. Preliminary studies done between March-2021 to May-2021 are not reported here.

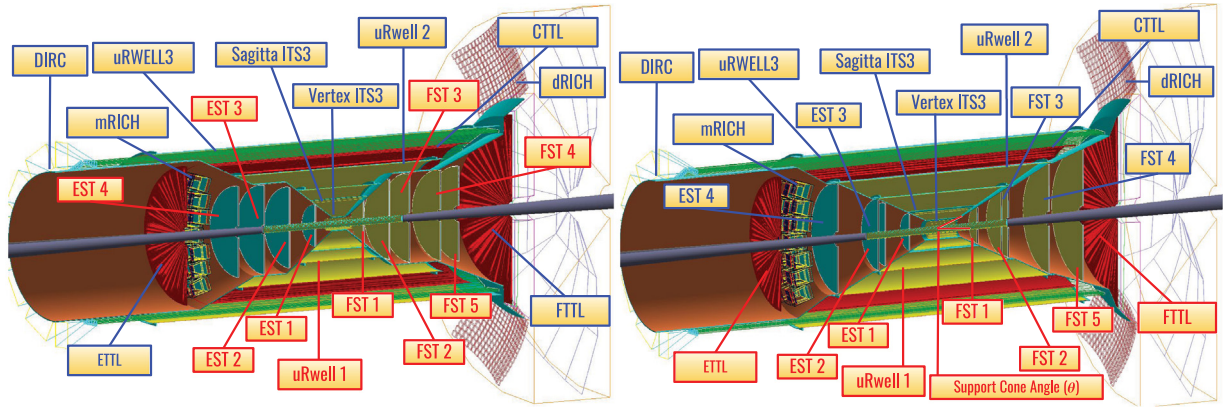


Fig. 5. Tracking and PID system in the non-projective (left) and the ongoing R&D projective (right) designs: the two figures show the different geometry and parametrization of the ECCE non-projective design (left) and of the ongoing R&D projective design to optimize the support structure (right). Labels in red indicate the sub-detector systems that were optimized, while the labels in blue are the sub-detector systems that were kept fixed due to geometrical constraint. The non-projective geometry (left) is a result of an optimization on the inner tracker layers (labeled in red) while keeping the support structure fixed. The angle made by the support structure to the IP is fixed at about 36.5° . The projective geometry (right) is the result of an ongoing project R&D to reduce the impact of readout and services on tracking resolution.

TOF detectors were modeled as simplified silicon disks at first; the outer trackers had more fine-grained simulations implemented, with realistic support structures and services implemented. The optimization pipelines included various combinations of detector technologies for the inner trackers. At the end of this phase, a decision on the choice of the barrel technology and the disk technologies was made using the AI results.

Phase II optimization. ⁶ These pipelines had a more realistic implementation of the support structure incorporating cabling, support carbon fiber, cooling system, etc. More detailed simulation of the PID Detectors (e.g., DIRC bars and dRICH sub-systems) were integrated as well as fine-grained simulations of TTL layers (CTTL, ETTL, FTTL) previously simulated as simple silicon layers modules. More stringent engineering constraints were considered such as the sensor size for MAPS detector (ITS3). This phase also considered an asymmetric arrangement of the detectors in the endcap regions, with a maximum of 4 EST disks in the electron-going end-cap and 5 FST disks in the hadron-going endcap: due to this asymmetric spatial arrangement, the angle subtended by detectors in the two endcap regions could be varied. This eventually developed into the idea of a projective geometry in a pipeline that characterizes an ongoing R&D project for optimizing the design of the support structure.

A detailed description of the most recent parametrization used for the detector proposal can be found in [Appendix A](#), along with the parametrization used in an ongoing R&D project to optimize the support structure of the inner tracker.

[Fig. 5](#) shows a comparison of the ECCE reference non-projective design and the projective design from the ongoing R&D, both of which resulted from the AI-assisted procedure described in this paper.

4.1. Encoding of Design criteria

Design criteria need to be encoded to steer the design during the optimization process. For each design point we need to compute the corresponding objectives f_m , namely the momentum resolution, angular resolution, and Kalman filter efficiency.

We will refer in the following only to the more recent Phase II optimization.⁷ Phase II has been characterized by two types of optimization pipelines: the first used a parametrization of the inner tracker during the optimization process and led to the ECCE tracker non-projective

Table 1

Summary of the hyperparameters of the design optimization: the values reported in the table have been used during the optimization of the non-projective design of the ECCE tracker. For completeness and when they differ from the non-projective case, we also report in parentheses the values corresponding to the ongoing R&D project for the projective design of the support structure.

Description	Symbol	Value
Population size	N	100
# objectives	M	3
Offspring	O	30
Design size	D	11 (9)
# calls (tot. budget)	–	200
# cores	–	same as offspring
# charged π tracks	N_{trk}	120k
# bins in η	N_η	5
# bins in p	N_p	10

design; the second branched off the first as an independent R&D effort that included the parametrization of the support structure and led to a projective design.

Details on the two types of optimization pipelines can be found in the following tables: [Table 1](#) describes the main hyperparameters and the dimensionality of the optimization problem, in particular of the design space and the objective space; [Table 2](#) reports the range of each design parameter⁸; [Table 3](#) summarizes the constraints for both the non-projective and projective geometries.

The parameters corresponding to the non-projective case in [Table 2](#) were optimized sequentially over at least three optimization pipelines, with each having a set of parameters fixed. For instance, the EST/FST disks were optimized during the first iteration assuming symmetric design, and for the subsequent iterations, the EST disks were optimized further in the electron endcap region. Two types of geometrical constraints have been implemented in the design optimization problem according to the definitions of [Table 3](#): (a) *strong constraints* cannot be violated during the optimization (e.g., a safe minimum distance between the disks of 10 cm has been considered) and hence result in high penalties during this process; (b) *soft constraints* can be violated by the MOO and then one can quantify the degree of unfeasibility (see [13]); soft constraints have been introduced to control the coverage of the disks and central layers with sensors, e.g., the difference between the outer and inner radii of each disk, namely $R_{\text{out}} - R_{\text{in}}$, is required to be a multiple of the sensor cell size (17.8 mm \times 30.0 mm); this requirement

⁶ Phase II corresponds to optimization pipelines that run from Sept-2021 to Nov-2021.

⁷ Similar considerations apply also for Phase I optimization.

⁸ The design points are normalized in the range [0-1], using a min-max scaler $x_i = (x_{\text{max}} - x_{\text{min}}) + x_{\text{min}}$, where x_i is the normalized design point with a un-normalized design point x generated between the range $[x_{\text{min}}, x_{\text{max}}]$.

Table 2

Ranges of the design parameters: the table summarizes the design points that are optimized for the non-projective and the projective case. The optimization range is also given for each parameter.

ECCE design (non-projective)	
Design parameter	Range
μ RWELL 1 (Inner) (r) Radius	[17.0, 51.0 cm]
μ RWELL 2 (Inner) (r) Radius	[18.0, 51.0 cm]
EST 4 z position	[-110.0, -50.0 cm]
EST 3 z position	[-110.0, -40.0 cm]
EST 2 z position	[-80.0, -30.0 cm]
EST 1 z position	[-50.0, -20.0 cm]
FST 1 z position	[20.0, 50.0 cm]
FST 2 z position	[30.0, 80.0 cm]
FST 3 z position	[40.0, 110.0 cm]
FST 4 z position	[50.0, 125.0 cm]
FST 5 z position	[60.0, 125.0 cm]
ECCE ongoing R&D (projective)	
Design parameter	Range
Angle (Support Cone)	[25.0°, 30.0°]
μ RWELL 1 (Inner) Radius	[25.0, 45.0 cm]
ETTL z position	[-171.0, -161.0 cm]
EST 2 z position	[45, 100 cm]
EST 1 z position	[35, 50 cm]
FST 1 z position	[35, 50 cm]
FST 2 z position	[45, 100 cm]
FST 5 z position	[100, 150 cm]
FTTL z position	[156, 183 cm]

Table 3

Constraints in the design optimization: the table summarizes constraints for both the projective and the non-projective designs. The reader can find a thorough explanation in the text.

Sub-detector	Constraint	Description
EST/FST disks	$\min \left\{ \sum_i^{disks} \left \frac{R'_{out} - R'_{in}}{d} - \left\lfloor \frac{R'_{out} - R'_{in}}{d} \right\rfloor \right \right\}$	soft constraint: sum of residuals in sensor coverage for disks; sensor dimensions: $d = 17.8$ (30.0) mm
EST/FST disks	$z_{n+1} - z_n \geq 10.0$ cm	strong constraint: minimum distance between 2 consecutive disks
sagitta layers	$\min \left\{ \left \frac{2\pi r_{sagitta}}{w} - \left\lfloor \frac{2\pi r_{sagitta}}{w} \right\rfloor \right \right\}$	soft constraint: residual in sensor coverage for every layer; sensor strip width: $w = 17.8$ mm
μ RWELL	$r_{n+1} - r_n \geq 5.0$ cm	strong constraint: minimum distance between μ Rwell barrel layers

can be slightly violated, since minor adjustments on tiling up the pixels can be done post hoc in such a way that the detector geometry is still realizable.

Two additional constraints are implicitly taken into account by internal parametrization and reflected in the GEANT4 design: the inner radii of the disks in the endcap and the inner vertex layer in the barrel have to be compatible with the beam envelope dimensions; the barrel layers lengths and the outer radii of the disks have to be compatible with the tracking support structure [25]. Also, potential overlaps among modules are checked before and during the optimization.

These constraints are common to non-projective and projective designs. For more details on the parametrizations and on the corresponding detector performance the reader can refer to [Appendix A](#) and [Appendix B](#), respectively.

The objectives depend on the kinematics and are calculated in 5 main bins in pseudorapidity (η): (i) $-3.5 \leq \eta < -2.0$ (corresponding to the electron-going direction), (ii) $-2.0 \leq \eta < -1.0$ (the transition region in electron-going direction), (iii) $-1 \leq \eta < 1$ (the central barrel), (iv) $1 \leq \eta < 2.0$ (the transition region in the hadron-going direction) and (v)

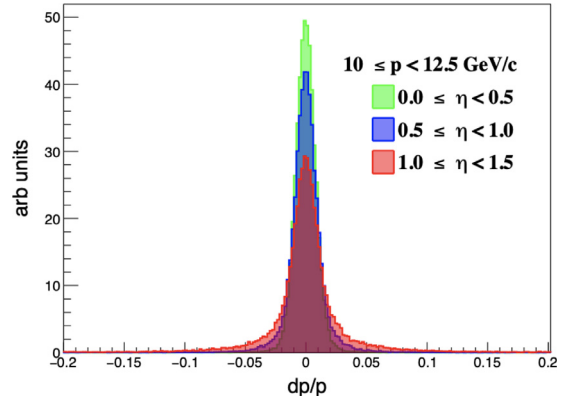


Fig. 6. Fit strategy: a double-Gaussian fit function is utilized to extract the resolutions. Such a fit function provided good reduced χ^2 and more stable extractions compared to single-Gaussian fits. The resolution is obtained as an average of the two σ 's weighted by the relative areas of the two Gaussians according to Eq. (3). The figure represents the results corresponding to a particular bin in η and p .

$2.0 \leq \eta < 3.5$ (the hadron-going direction). The rationale behind this binning is a combination of different aspects: the correspondence with the binning in the EIC Yellow Report [1], the asymmetric arrangement of detectors in electron-going and hadron-going directions, and the division in pseudorapidity between the barrel region and the endcap. Particular attention is given to the transition region between barrel and endcaps as well as at large $|\eta| \sim 3.5$ close to the beamline.

Charged pions are generated uniformly in the phase-space that covers the range in momentum magnitude $p \in [0, 20]$ GeV/c and the range in pseudorapidity $\eta \in (-3.5, 3.5)$. Each bin in η is further subdivided into 15 bins in momentum p . For each design point, we simulated $N_{trk} = 120k$ charged pions.⁹ This number ensured large enough statistics over the entire phase space and the stability of the fits in all of the bins of Eqs. (4).

In order to calculate the relative momentum (cf. [Fig. B.13](#)) and absolute angular resolution (cf. [Fig. B.14](#)) we fit the following objectives:

$$\begin{aligned} \frac{\Delta p}{p} &= \frac{p_{reco} - p_{true}}{p_{true}} \\ \Delta\theta &= \theta_{reco} - \theta_{true} \end{aligned} \quad (2)$$

Following the definitions of Eq. (2), histograms of the relative momentum resolution and the absolute angular resolution are produced for each bin in η and p , and the corresponding fits are calculated. Using single-Gaussian (SG) fits (also utilized in the Yellow Report [1]) implies systematically better resolutions but worse reduced χ^2 : therefore we decided to utilize double-Gaussian (DG) fits, as shown in [Fig. 6](#). This provided a more robust fit strategy. The reduced χ^2 range with DG fits ranges from 1.2 to 2.8 at most, with the majority of the fits stable at lower values. The largest numbers correspond either to the transition between the barrel and endcaps – where tracks cross more material in the non-projective design – or to large pseudorapidity, particularly close to the inner radii of the disks. By using SG fits, the reduced χ^2 values can be as large as 10–20 in the transition region. A detailed study comparing SG to DG fits is shown in [Fig. B.13](#).

⁹ From Phase I to Phase II, the design became asymmetric in the two endcaps, therefore we needed to extend the η -coverage and increase the statistics. The momentum range was reduced to $[0, 20]$ GeV/c to optimize the computing budget.

The final DG resolution has been defined as an average of the two σ 's weighted by the relative areas of the two Gaussians:¹⁰

$$\sigma_{DG} \sim \frac{A_1 \sigma_1 + A_2 \sigma_2}{A_1 + A_2} \quad (3)$$

The results obtained for the resolutions in each bin corresponding to each new design point are divided by the values corresponding to the baseline design, so that in each bin a ratio $R(f)_{p,\eta}$ is provided. Finally a weighted sum of these ratios is performed to build a global figure of merit (for both the relative momentum and the angular resolutions):

$$R(f) = \frac{1}{N_\eta} \sum_\eta \left(\frac{\sum_p w_{p,\eta} \cdot R(f)_{p,\eta}}{\sum_p w_{p,\eta}} \right) \quad (4)$$

where the objective function f is either the momentum or the angular resolution described by Eq. (2), and the weight $w_{p,\eta} = 1/\sigma^2(f_{p,\eta})$ is calculated in each η, p bin and it is proportional to the inverse of the variance corresponding to the objective functions f .

An additional objective function has been included in the optimization problem: this is a global objective function corresponding to the fraction of tracks that are not reconstructed by the Kalman filter (KF [26]), or equivalently the KF inefficiency:

$$R(KF) = \frac{N(\text{tracks not reconstructed})}{N(\text{total number of tracks})} \quad (5)$$

Following the above definitions, the design problem corresponds to minimizing the objective functions defined in Eqs. (4), (5): a weighted average ratio smaller than one and a KF inefficiency as close as possible to zero correspond to an improvement in performance compared to the baseline design.

4.2. Convergence and performance at Pareto front

We remind the reader that the Pareto front is the set of trade-off solutions to our problem. Fig. 7 shows the convergence plot obtained utilizing the hypervolume, which is a standard metric [27] used as a performance indicator for the optimization problem. Given a reference point r as outlined in Fig. 2, this metric calculates the hypervolume subtended by the approximated Pareto front in the objective space. The MOO steers the approximated front towards the true front, therefore the growth of the hypervolume is monotonic by construction and in an ideal case it asymptotically converges to the true Pareto front.

One can deploy a variety of convergence criteria and analyze the convergence as mentioned in [13].¹¹ For our studies, and given the complexity of our design problem summarized in Table 1, the termination criterion is defined using a moving window of 5 consecutive calls and is triggered if either of the following conditions are true: the relative change in any of the design parameters is less than 10^{-5} (*design tolerance*); the relative improvement in the constraint violation is less than 10^{-7} (*constraint tolerance*); the improvement in any of the objectives is less than $2.5 \cdot 10^{-3}$ (*objective tolerance*). Finally we also utilize a total budget of 200 calls, after which we interrupt the execution of MOO.

A petal diagram is used to visualize the values of three objectives corresponding to one of the solutions extracted from the Pareto front. Checkpoints are created to store the NSGA-II-updated population of design points. A survey of the detector performance is created after each call to monitor the potential anomaly behavior of the fits. The fitting procedure is quite stable: if an exception occurs the analysis has been automated to adjust the fitting parameters and ranges. In case of persistent anomalous behavior, a flag is raised, the critical design point purged from the population, and examined.

¹⁰ A different definition could be based on the weighted average of the variances $\sigma_{1,2}^2$ to obtain the final variance σ_{DG}^2 . This typically implied a few % relative differences on the final value of σ_{DG} which has been considered a negligible effect.

¹¹ Early stopping can occur if a change in the hypervolume is within the tolerance observed after a certain number of evaluations.

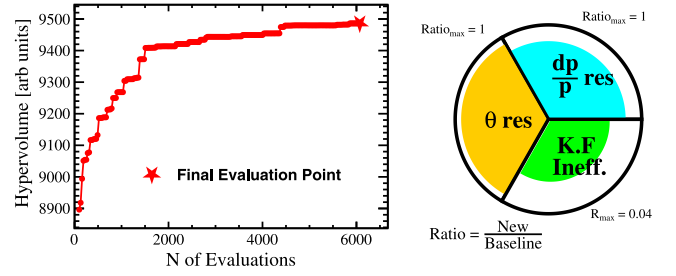


Fig. 7. Metric for convergence and performance at Pareto front: (left) The hypervolume can be used as a metric for convergence. Checkpoints are created during the optimization and snapshots of the evolving designs are saved. (right) A petal diagram with the three objectives corresponding to one solution in the Pareto front. The momentum and angular resolutions are expressed as ratios with respect to a baseline design to improve; the KF inefficiency is taken as an absolute value. An optimal design optimizes all of the above-defined objectives.

The improvement obtained with the continual multi-objective optimization process is summarized in Fig. 8, where the momentum resolution obtained during phase-I optimization using a preliminary detector concept is compared to both the non-projective and the projective R&D designs which are instead derived from fully developed simulations in phase-II optimization.

A detailed description of the optimized performance for all the objectives (momentum, angular resolutions and Kalman Filter efficiency) can be found in Appendix B.

4.3. Physics analysis

To show a comparison in physics performance between the non-projective and projective designs, we analyzed D^0 meson decay into $\pi^+ K^-$. Data have been produced utilizing SIDIS events generated with Pythia6 [28], corresponding to ep events with $18 \text{ GeV} \times 275 \text{ GeV}$ and high Q^2 .¹²

In Fig. 9, the $\pi^+ K^-$ invariant mass is fitted with a double-sided Crystal Ball function [30,31].

The decay events are selected in such a way to have at least one particle (either π or K , or both) in the pseudorapidity bin $1.0 < |\eta| < 1.5$, where the projective design is expected to improve the performance by concentrating all the material in a smaller dead area compared to the non-projective design.

The analysis shows that the resolution obtained with the projective design is improved by more than 10% relative to that obtained with the non-projective design. We also calculate the efficiency, defined as the number of reconstructed D^0 mesons divided by the number of true D^0 mesons. The efficiency obtained with the two designs is consistent within the statistical uncertainties.

5. Computing resources

Parallelization. A two-level parallelization has been implemented in the MOO framework: the first level creates the parallel simulations of design points, the second level parallelizes each design point (see Fig. 10). The evaluation itself can be distributed to several workers or a whole cluster with libraries like Dask [32].

Computing budget. Computing time studies have been carried out to evaluate the simulation time of each single design point as a function of the number of tracks generated. We made this study with simulations that included the tracking system and the PID system and estimated an effective simulation time of 0.2 s/track after removing an initial latency

¹² Production: prop.5/prop.5.1; generator: pythia6; kinematics: ep-18 \times 275-q2-high. More info can be found in [29].

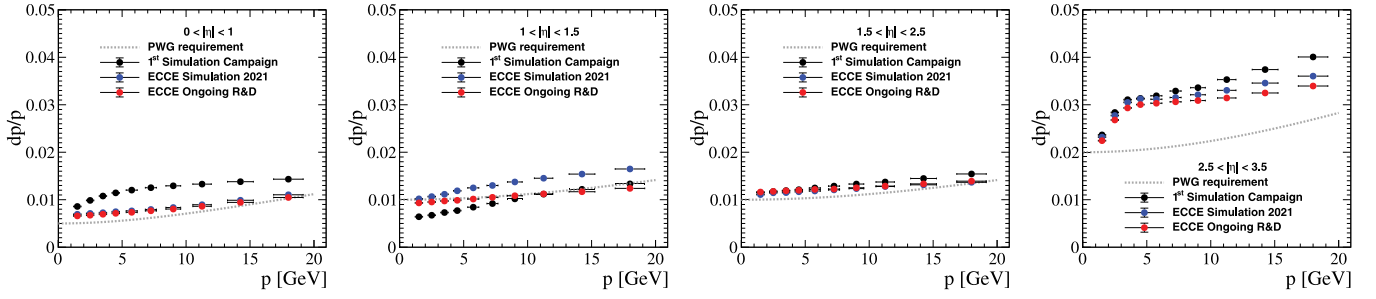


Fig. 8. Evolution of momentum resolution driven by the continued MOO: Momentum resolutions in four bins of η . From left to right, respectively: $|\eta| < 1.0$ corresponding to the barrel, $1.0 < |\eta| < 1.5$ corresponding to the transition region, and $1.5 < |\eta| < 2.5$ and $2.5 < |\eta| < 3.5$ corresponding to the two end-caps. Black points represent the first simulation campaign, and a preliminary detector concept in phase-I optimization which did not have a developed support structure; blue points represent the fully developed simulations for the final ECCE detector proposal concept; red points the ongoing R&D for the optimization of the support structure. Compared to black, there is an improvement in performance in all η bins with the exception of the transition region, an artifact that depends on the fact that black points do not include a realistic simulation of the material budget in the transition region. In the transition region, it can be also appreciate the improvement provided by the projective design.

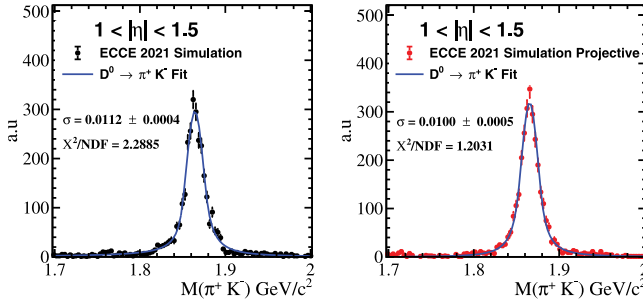


Fig. 9. Analysis of SIDIS events with AI-assisted detector designs: The $\pi^+ K^-$ invariant mass obtained from the SIDIS events reconstructed in ECCE. A comparison of the resolution on the D^0 meson obtained with the non-projective and the projective design of the ECCE tracker is shown. In order to emphasize the improvement in the region between $1.0 < |\eta| < 1.5$, a selection was made such that both the tracks π^+ or the K^- are within the region of interest. The resolution with the projective design is improved by more than 10% relative to the resolution obtained with the non-projective design.

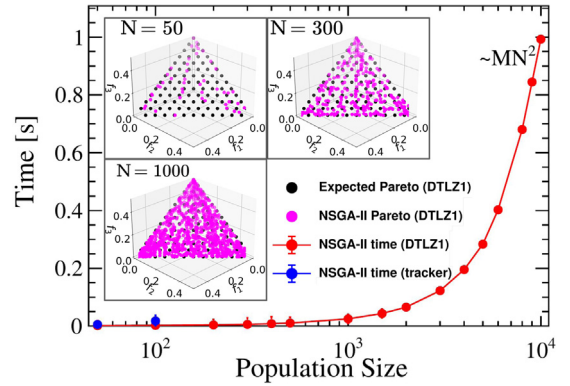


Fig. 11. The computing time taken by the genetic algorithm and the sorting in NSGA-II: Performance has been benchmarked with test problems like DTLZ1 [17] (red points) and the scaling $\sim MN^2$ (see Table 1) has been verified with convergence to the Pareto front. The simulation time dominates the AI time during the optimization process (blue point). A two-level parallelization has been introduced in the framework to reduce this bottleneck. The AI contribution typically becomes dominant when a very large population size is needed to accurately approximate the Pareto front (cf.[34]).

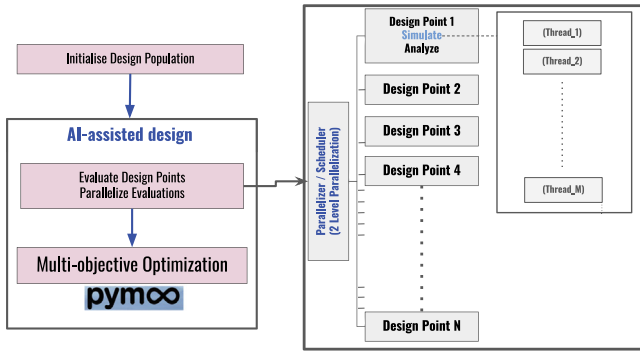


Fig. 10. Flowchart of the two-level parallelization implemented for detector design: Multiple design points are produced in parallel, and each design point is parallelized. The design is assisted by an AI-strategy, in this case MOO with evolutionary algorithms supported by pymoo [13].

time. Similarly we made studies of the computing time taken by the AI-based algorithm in generating a new population of design points. Results of these studies are summarized in Fig. 11.

A larger population allows to approximate the Pareto front with larger accuracy. Extension of the design parameter space and the objective space to larger dimensionality implies a larger amount of CPU time which is mainly dominated by simulations if the population size remains smaller than 10^4 – 10^5 , see Fig. 11.

For our goals the optimization pipelines of the ECCE tracking system were parametrized with 10–20 design parameters and 3–4 objectives;

this allowed us to achieve good convergence with evolutionary MOO using a two-level parallelization strategy, and deployment on single nodes of 128 CPU cores available on the sci-comp farm at Jefferson Lab [33].¹³ The population size N is chosen considering the dimensionality of the design space and objective space, the simulation time of each design point, and the computing resources available.

Planned activities. As described in this document, detector optimization with AI is an essential part of the R&D and design process and it is anticipated to continue after the detector proposal. The AI-assisted design optimization of the ECCE inner tracker was based on evolutionary algorithms. During the detector proposal multiple optimization pipelines were run each with a population size of 100, representing different detector design configurations. At each iteration, AI updated the population. The total computing budget for an individual pipeline amounted to approximately 10k CPU-core hours. This number depends on the dimensionality of the problem. Larger populations may need to be simulated to cope with the increased complexity in order to improve the accuracy of the approximated Pareto front. Different AI-based strategies will be compared.

Activities are planned to continue the detector optimization: new optimization pipelines can deal with a larger parameter space to include a system of sub-detectors or to combine tracking and PID in the

¹³ Work is in progress to efficiently distribute the optimization pipeline to multiple nodes.

optimization process. We also plan to optimize other sub-detectors like, e.g., the dRICH, leveraging on the expertise internal to the ECCE collaboration regarding specifically the design of the dRICH with AI-based techniques [11]. As a future activity we aim to encode physics-driven objectives in the MOO problem. A thorough comparison of results obtained with different AI-based strategies (e.g., MOO based on genetic algorithms or Bayesian approaches) can be also studied.

We anticipate for 2022 roughly 1M CPU-core hours for these activities.

6. Summary

Large scale experiments in high energy nuclear physics entail unprecedented computational challenges and the optimization of their complex detector systems can benefit from AI-based strategies [6].

In this paper we described the successful implementation of a multi-objective optimization approach to steer the multidimensional design of the ECCE tracking system, taking into account the constraints from the global detector design. This work was accomplished during the EIC detector proposal, and was characterized by a continued optimization process where multiple optimization pipelines integrating different configurations of sub-detectors were compared using full GEANT4 simulations. The insights provided by AI in such a multi-dimensional objective space characterizing the detector performance (e.g., tracking efficiency, momentum and angular resolutions), combined to other aspects like risk mitigation and costs reduction, helped selecting the candidate technology of the ECCE tracker. This approach is being used in an ongoing R&D project where the design parametrization has been extended to include the support structure of the tracking system.

The design optimization can be also extended to tune the parameters of a larger system of sub-detectors. Physics analyses are at the moment done after the optimization for a given detector design solution candidate, but they can be encoded during the optimization process as physics-driven objectives in addition to objectives representing the detector performance.

Detector optimization with AI is anticipated to continue after the detector proposal, and activities are planned to further optimize the tracking system, including PID sub-detectors, particularly the dual-RICH [11].

Acknowledgments

We thank the EIC Silicon Consortium for cost estimate methodologies concerning silicon tracking systems, technical discussions, and comments. We acknowledge the important prior work of projects eRD16, eRD18, and eRD25 concerning research and development of MAPS silicon tracking technologies.

We thank the EIC LGAD Consortium for technical discussions and acknowledge the prior work of project eRD112.

We thank (list of individuals who are not coauthors) for their useful discussions, advice, and comments.

We acknowledge support from the Office of Nuclear Physics in the Office of Science in the Department of Energy, USA, the National Science Foundation, USA, and the Los Alamos National Laboratory Laboratory Directed Research and Development (LDRD), USA 20200022DR.

Appendix A. Details on parametrization

Tracking system parametrization

Vertex layers. There are three vertex barrel layers in the ECCE tracking system made of MAPS technology. The vertex cylinder consists of strips which are made of pixels, where the individual sensor unit cell size is 17.8 mm × 30.0 mm. The length of the vertex layers is fixed at 27 cm;

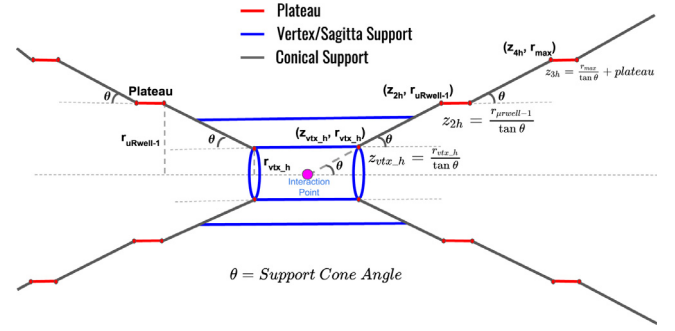


Fig. A.12. Parametrization of the support cone for the inner tracker: The inner tracker support is characterized by five variables: θ (the angle of projection of the support cone structure), r_{vertex} (radius of vertex support structure), $r_{\mu\text{Rwell-1}}$ $\mu\text{Rwell-1}$ radius, plateau length, r_{max} maximum allowed radius of inner tracker.

the radii of the three vertex layers are fixed to 3.4, 5.67, 7.93 cm, respectively. For the non-projective design, the angle of the support structure with respect to the interaction point is fixed ($\theta = 36.5^\circ$) and the radius of the support is at 6.3 cm, while the length of it is 17 cm. For the projective design, the radius of the support structure is the same, while the length is calculated based on the angle of projection and the radius as shown in Fig. A.12.

Sagitta layers. There are two sagitta barrel layers in the ECCE tracking system. The sagitta barrel layers are made of MAPS technology and have fixed length of 54 cm. For the non-projective design the radii of the sagitta layers are 21.0, 22.68 cm, respectively. For the projective parametrization, the radius of the sagitta barrel is calculated such that there are no gaps in the acceptance of the region enclosed by the barrels, according to the following equation: $r_{\text{sagitta}} = \frac{l_{\text{sagitta}}}{2} \tan \theta$.

The radius of the sagitta layers is also constrained since the strips have fixed width $w = 17.8$ mm; therefore we want to minimize the quantity:

$$\min \left\{ \left| \frac{2\pi r_{\text{sagitta}}}{w} - \left\lceil \frac{2\pi r_{\text{sagitta}}}{w} \right\rceil \right| \right\},$$

where $\lceil x \rceil$ represents the ceiling of x .

μRwell layers. In the ECCE tracking system there are three cylindrical μRwell layers, each endowed with a support ring. An extended supporting plateau is included at either ends of the μRwell to rest the entire cylindrical detector on this platform. This results in a constant shift of the support cone by the plateau length (5 cm) as shown in Fig. A.12. For both the non-projective design and the projective design the $\mu\text{Rwell-1}$ radius is a free parameter. The length of the $\mu\text{Rwell-1}$ is calculated based on the angle of the conical support structure. In the non-projective design we have the conical support structure angle fixed ($\theta = 36.5^\circ$), therefore the length of $\mu\text{Rwell-1}$ depends only on its radius; $\mu\text{Rwell-2}$ has its radius as a free parameter; since the angle of the conical support structure is fixed the length of $\mu\text{Rwell-2}$ depends on its radius. In the projective design instead the $\mu\text{Rwell-2}$ has a fixed radius of 51 cm (i.e., $r_{\text{max}} + 1$ cm). The length of the $\mu\text{Rwell-2}$ is calculated based on the angle of the conical support structure. The length of the μRwell takes into account the constant shift due to the plateau. The dimensions of $\mu\text{Rwell-3}$ are fixed in both non-projective and projective designs; the $\mu\text{Rwell-3}$ is outside of the inner tracking system and it has radius of 77 cm and a total length is 290 cm.

EST/FST disks. For both the non-projective and projective designs, R_{min} of the disks must be compatible with the beam pipe envelope which increases in radius as a function of z ; R_{out} of the disks is parametrized to be compatible with the support cone structure shown in Fig. A.12 which has an angle θ that is variable in the projective design and fixed in the non-projective case. For the non-projective

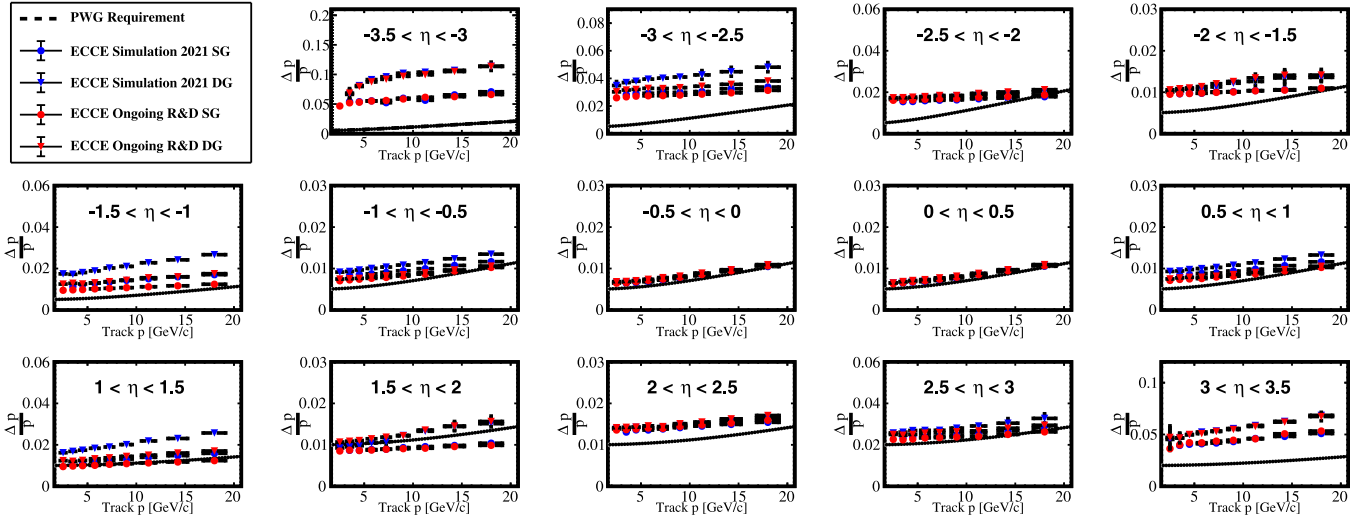


Fig. B.13. Momentum resolution for the ECCE tracker reference design and the projective design (ongoing R&D): The $\Delta p/p$ are extracted with single-Gaussian (SG) fits (circles) to be consistent with the YR PWG and compared to double-Gaussian (DG) fits (triangles). DG (see Eq. (33)) is used in the AI-assisted pipelines as it provides more stable fits. DG results are systematically larger than the SG ones as they capture more realistic distributions (see Fig. 6). This effect is larger in $1 < |\eta| < 1.5$ for the non-projective design (blue points) since more material is traversed by the tracks; the same effect is moderately present also in $0.5 < |\eta| < 1$ for the same reason. In the same bins, the ongoing R&D design shows a significant reduction of the impact of the readout and services and SG (red circle) and DG (red triangle) are close to each other. DG points are systematically larger in $3 < \eta < 3.5$ and in $-3.5 < \eta < -3.0$, and this corresponds to edge effects at large $|\eta|$ values close to the inner radii of the disks in the endcaps. In the region $-1 < \eta < 3.0$ $\Delta p/p$ is consistent with the YR physics requirements with SG. Note that physics in the backward region relies on the EM calorimeter also, and in the proposal it has been demonstrated that larger resolutions are acceptable.

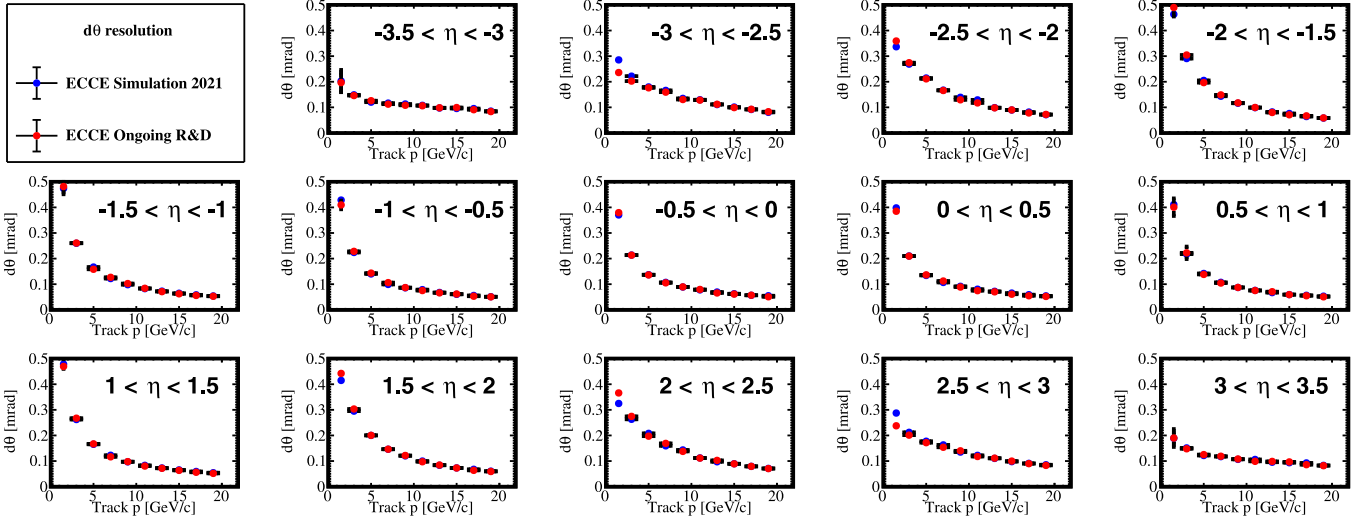


Fig. B.14. Angular resolution ($d\theta$) for ECCE projective geometry and ECCE simulation: The resolution shows an improvement in the transition region between $1.0 < |\eta| < 1.5$. These plots have been produced using double-Gaussian fits.

design, the z positions of the disks were all free parameters in the first optimization pipelines. However, to maximize the hit efficiency, some disks have been eventually placed within the support cone at the beginning of every plateau (Fig. A.12 with fixed angle $\theta = 36.5^\circ$). Therefore, two disks in the electron-going direction and two disks in the hadron-going direction are not free to vary in z . For instance, consider Fig. 5 (right), where EST3, EST4, FST3, FST4 are placed at the begin of the plateau, whereas the disks EST1, EST2, FST1, FST2, FST5 are free to vary in z position. The same parametrization is extended to the projective design and made compatible with a varying conical support structure.

As the disks are tiled up using MAPS pixels, the difference between R_{min} and R_{out} is constrained to optimize the sensor coverage for all disks; this is implemented by means of two functions, namely:

$$\min \left\{ \sum_i^{all \text{ disks}} \left| \frac{R_{out}^i - R_{in}^i}{d} - \left\lfloor \frac{R_{out}^i - R_{in}^i}{d} \right\rfloor \right| \right\},$$

where $d = 17.8$, and 30.0 mm. This limits the amount of violation made by a design solution.

TOF system. The central barrel TOF (CTTL) is an AC-LGAD based TOF detector with a fixed radius of 64 cm and a fixed length of 280 cm. The TOFs at the electron-going endcap (ETTL) and the hadron-going endcap (FTLL) are AC-LGAD-based TOF disks. For the non-projective design the TOF detectors have fixed dimensions. For the projective design the TOF detectors in the end cap regions have their z positions as free parameters. R_{in} and R_{out} of the ETTL/FTLL disks depends on the position of the disk z . The R_{in} of the disk should be compatible with the radius of the beam envelope which increases linearly as a function of z ; R_{out} of the disks varies as a function of z such that the acceptance coverage by the ETTL/FTLL is roughly unaltered.

PID detectors. The Detection for Internally Reflected Cherenkov light (DIRC) is a detector for PID in the barrel region. DIRC system has

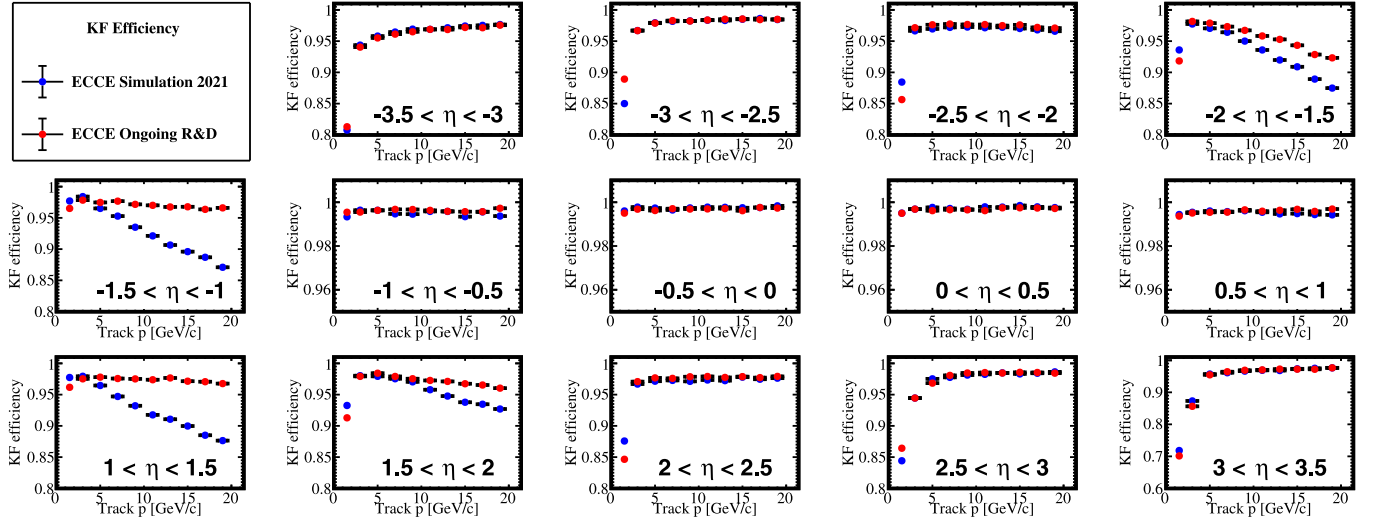


Fig. B.15. KF Efficiency for ECCE projective geometry and ECCE simulation: Reconstruction efficiency shows an improvement in the transition region between $1.0 < |\eta| < 1.5$. There is a significant drop in the transition region caused due to multiple scattering of the support structure.

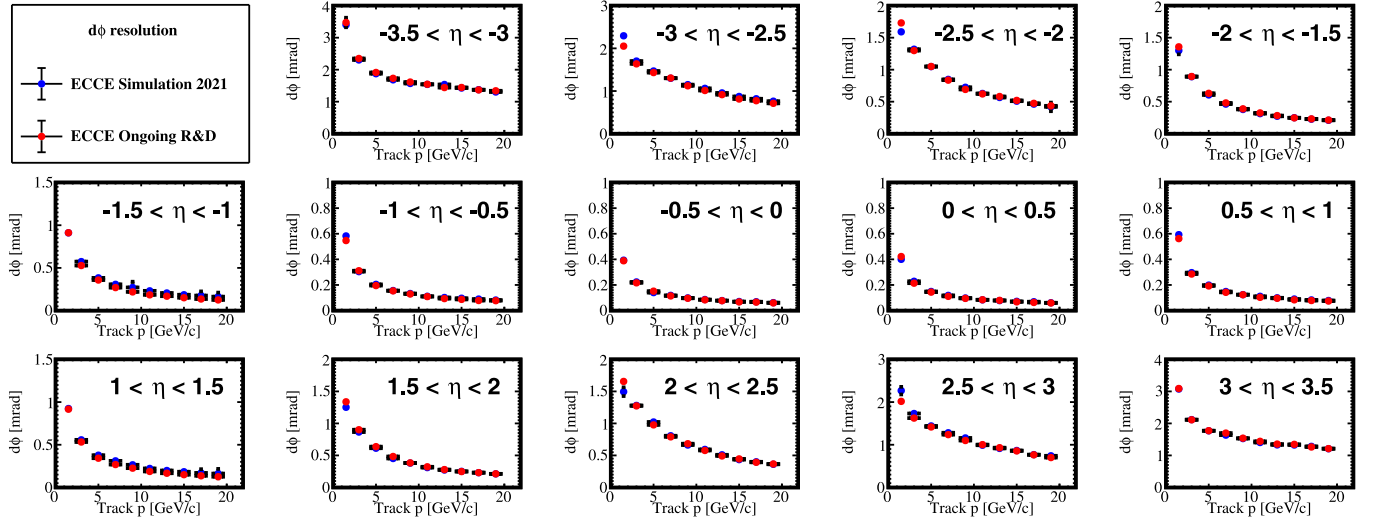


Fig. B.16. Angular resolution ($d\phi$) for ECCE projective geometry and ECCE simulation: The resolution shows an improvement in the transition region between $1.0 < |\eta| < 1.5$. These plots have been produced using double-Gaussian fits.

fixed dimensions and occupies a radial space from 71.5 cm to 76.6 cm. The modular RICH (mRICH) is a Ring Imaging Cherenkov detector system in the e-going direction with fixed dimensions. mRICH has a z position starting at -135 cm extending in z to -161 cm. The dual Radiator Imaging Cherenkov (dRICH) detector is a detector system in the forward direction with fixed dimensions. dRICH has a z position starting at 180 cm and extends up to 280 cm.

The thickness of the detectors and support structures are also taken into account to avoid overlaps between the detectors. The most recent optimization pipelines were extended to also include in the parametrization the outer tracking layers in the two endcaps, as explained in Section 4.1. An overlap check is performed each time a new design point is evaluated during the optimization process.

Support structure parametrization

The implementation of the projective geometry of the inner tracker is described in Fig. A.12, which shows the parametrization used for the support cone structure of the inner tracker. Some parameters have been considered fixed and other free to vary within their ranges. Parameters that are fixed typically do not have much room for optimization

considering the constraints of the design and potential overlaps. The non-projective design can be realized by fixing the support structure angle to ($\theta = 36.5^\circ$) shown in A.12. Therefore, the non-projective design solutions are a subset of solutions that can be achieved by this parametrization.

Appendix B. Baseline and r&d designs

Resolutions and efficiency

A thorough comparison between the non-projective ECCE simulation and the ongoing R&D was carried out to optimize the support structure through a projective design. Fig. B.13 shows a study of the obtained momentum resolution in bins of pseudorapidity. Similarly Fig. B.14 shows the angular resolution while Fig. B.15 shows the Kalman filter efficiency as defined in Eq. (5).

Validation

Validation is performed by looking at figures of merit that are not used during the optimization process. In Section 4.3 we already

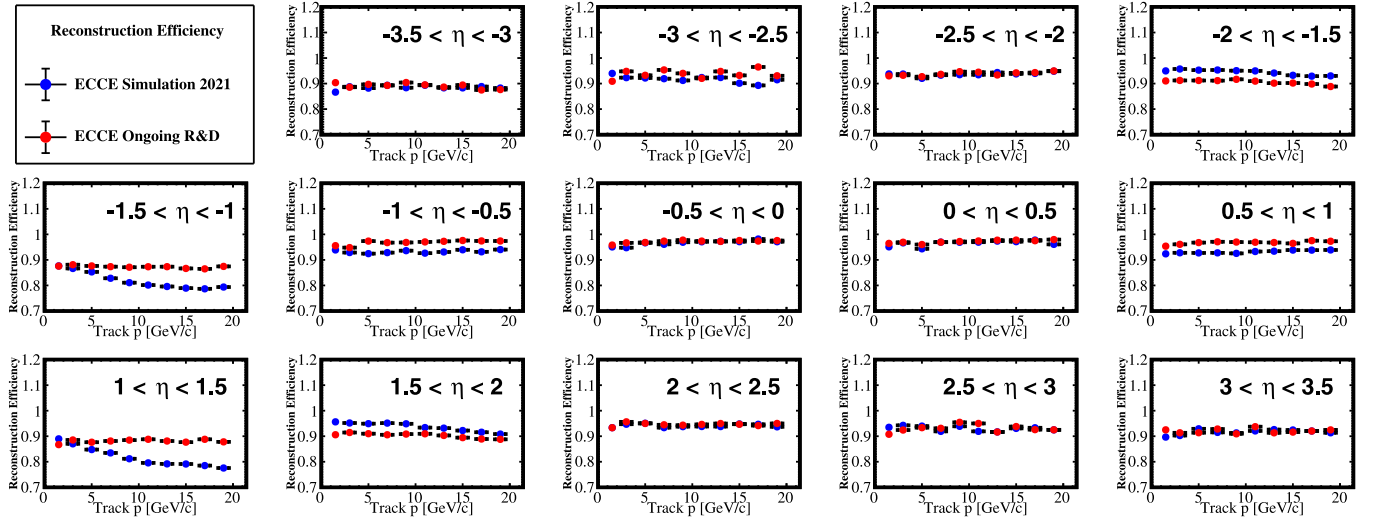


Fig. B.17. Reconstruction efficiency for ECCE projective geometry and ECCE simulation: Reconstruction efficiency shows an improvement in the transition region between $1.0 < |\eta| < 1.5$.

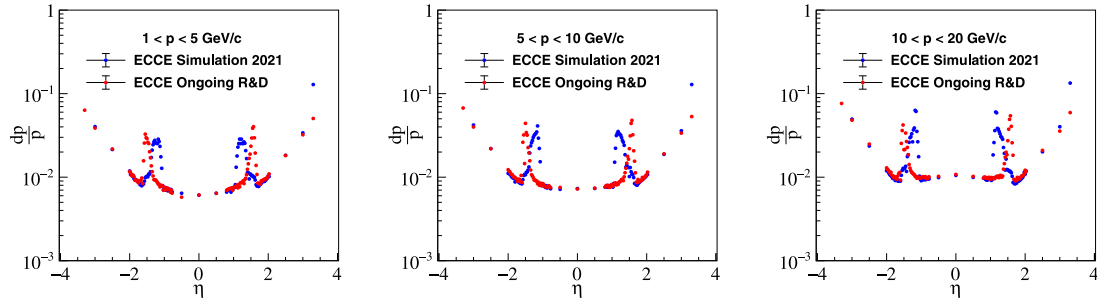


Fig. B.18. Comparison between non-projective and projective inner tracker: a comparison in momentum resolution as a function of the pseudorapidity η between the non-projective and the projective designs in three bins of momentum. The projective design concentrates the material in a smaller dead area resulting in better resolution on a wider range of the pseudorapidity.

described a physics analysis with SIDIS events that further consolidates our conclusions. We show here additional examples of validation: Fig. B.16 and Fig. B.17 display the azimuthal angular resolution and the reconstruction efficiency obtained for both the non-projective and the projective tracker designs. The azimuthal resolution looks consistent within the uncertainty while the reconstruction efficiency looks in general better for the projective design, particularly in the $1 < |\eta| < 1.5$ region where the non-projective design has a larger dead area that affects the reconstruction of tracks.

A comparison between the non-projective and projective designs of the inner tracker is also shown in Fig. B.18, where the projective design concentrates the material in a smaller dead area resulting in better resolution on a wider range of the pseudorapidity.

References

- [1] R.A. Khalek, A. Accardi, J. Adam, D. Adamiak, W. Akers, M. Albaladejo, A. Al-Bataineh, M. Alexeev, F. Ameli, P. Antonioli, et al., Science requirements and detector concepts for the electron-ion collider: EIC yellow report, 2021, arXiv preprint [arXiv:2103.05419](https://arxiv.org/abs/2103.05419), URL <https://arxiv.org/abs/2103.05419>.
- [2] The ECCE consortium, 2021, URL <https://www.ecce-eic.org>.
- [3] J.K. Adkins, et al., ECCE Consortium Collaboration, Design of the ECCE detector for the electron ion collider, 2022, arXiv preprint [arXiv:2209.02580](https://arxiv.org/abs/2209.02580), URL <https://arxiv.org/abs/2209.02580>.
- [4] R. Ent, EIC overview and schedule, 2021, The AI4EIC Workshop - First workshop on Artificial Intelligence for the Electron Ion Collider, <http://eic.ai>, URL https://indico.bnl.gov/event/10699/contributions/53658/attachments/36955/60868/AI4EIC_EIC_Overview_and_Schedule_090721_post.pptx.
- [5] S. Agostinelli, J. Allison, K.A. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, et al., GEANT4—A simulation toolkit, Nucl. Instrum. Methods Phys. Res. A 506 (3) (2003) 250–303.
- [6] C. Fanelli, Design of detectors at the electron ion collider with artificial intelligence, 2022, arXiv preprint [arXiv:2203.04530](https://arxiv.org/abs/2203.04530).
- [7] Z. Zhou, S. Kearnes, L. Li, R.N. Zare, P. Riley, Optimization of molecules via deep reinforcement learning, Sci. Rep. 9 (1) (2019) 1–10, <http://dx.doi.org/10.1038/s41598-019-47148-x>.
- [8] K. Guo, Z. Yang, C.-H. Yu, M.J. Buehler, Artificial intelligence and machine learning in design of mechanical materials, Mater. Horiz. 8 (4) (2021) 1153–1172, <http://dx.doi.org/10.1039/D0MH01451F>.
- [9] B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, Science 361 (6400) (2018) 360–365, <http://dx.doi.org/10.1126/science.aat2663>.
- [10] ATLAS collaboration, et al., AtlFast3: The next generation of fast simulation in ATLAS, 2021, arXiv preprint [arXiv:2109.02551](https://arxiv.org/abs/2109.02551).
- [11] E. Cisbani, A. Del Dotto, C. Fanelli, M. Williams, et al., AI-optimized detector design for the future Electron-Ion Collider: The dual-radiator RICH case, J. Instrum. 15 (05) (2020) P05009.
- [12] G. Debreu, Valuation equilibrium and Pareto optimum, Proc. Natl. Acad. Sci. USA 40 (7) (1954) 588.
- [13] J. Blank, K. Deb, pymoo: Multi-objective Optimization in Python, IEEE Access 8 (2020) 89497–89509.
- [14] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE Trans. Evol. Comput. 6 (2) (2002) 182–197.
- [15] D. Whitley, A genetic algorithm tutorial, Stat. Comput. 4 (2) (1994) 65–85.
- [16] K. Deb, A. Pratap, S. Moitra, Mechanical component design for multiple objectives using elitist non-dominated sorting GA, in: International Conference on Parallel Problem Solving from Nature, Springer, 2000, pp. 859–868.
- [17] H. Ishibuchi, R. Imada, Y. Setoguchi, Y. Nojima, Performance comparison of NSGA-II and NSGA-III on various many-objective test problems, in: 2016 IEEE Congress on Evolutionary Computation, CEC, IEEE, 2016, pp. 3045–3052.
- [18] S. Agostinelli, et al., GEANT4: A simulation toolkit, Nucl. Instrum. Meth. A506 (2003) 250–303, [http://dx.doi.org/10.1016/S0168-9002\(03\)01368-8](http://dx.doi.org/10.1016/S0168-9002(03)01368-8).
- [19] Eic, fun4all_coresoftware, URL https://github.com/eic/fun4all_coresoftware.

- [20] J.C. Bernauer, et al., Scientific Computing Plan for the ECCE Detector at the Electron Ion Collider, 2022, arXiv preprint [arXiv:2205.08607](https://arxiv.org/abs/2205.08607).
- [21] ECCE Consortium, Design and simulated performance of tracking systems for the ECCE detector at the electron ion collider, 2021, URL <https://www.ecce-eic.org/ecce-internal-notes>. ecce-note-det-2021-03.
- [22] G. Aglieri Rinella, et al., First demonstration of in-beam performance of bent monolithic active pixel sensors, 2021, [arXiv:2105.13000](https://arxiv.org/abs/2105.13000).
- [23] D. Colella, ALICE ITS 3: The first truly cylindrical inner tracker, 2021, [arXiv:2111.09689](https://arxiv.org/abs/2111.09689).
- [24] L. Wei, et al., Precision timing silicon detectors for a combined PID and tracking system at EIC, 2021, URL https://wiki.bnl.gov/conferences/images/0/01/ERD29_progress_report_March2021.pdf.
- [25] J. Arrington, R. Cruz-Torres, W. DeGraw, X. Dong, L. Greiner, S. Heppelmann, B. Jacak, Y. Ji, M. Kelsey, S.R. Klein, et al., EIC physics from an all-silicon tracking Detector, 2021, arXiv preprint [arXiv:2102.08337](https://arxiv.org/abs/2102.08337).
- [26] R. Fruhwirth, Application of Kalman filtering to track and vertex fitting, Nucl. Instrum. Methods Phys. Res. A 262 (2) (1987) 444–450, [http://dx.doi.org/10.1016/0168-9002\(87\)90887-4](https://doi.org/10.1016/0168-9002(87)90887-4), URL <http://www.sciencedirect.com/science/article/pii/0168900287908874>.
- [27] C. Fonseca, L. Paquete, M. Lopez-Ibanez, An improved dimension-sweep algorithm for the hypervolume indicator, in: 2006 IEEE International Conference on Evolutionary Computation, 2006, pp. 1157–1163, [http://dx.doi.org/10.1109/CEC.2006.1688440](https://doi.org/10.1109/CEC.2006.1688440).
- [28] T. Sjöstrand, S. Mrenna, P. Skands, PYTHIA 6.4 physics and manual, J. High Energy Phys. 05 (2006) 026, [http://dx.doi.org/10.1088/1126-6708/2006/05/026](https://doi.org/10.1088/1126-6708/2006/05/026), [arXiv:hep-ph/0603175](https://arxiv.org/abs/hep-ph/0603175).
- [29] ECCE simulation working group page, URL https://wiki.bnl.gov/eicug/index.php/ECCE_Simulations_Working_Group.
- [30] M. Aaboud, et al., Search for resonances in diphoton events at $\sqrt{s}=13$ TeV with the ATLAS detector, J. High Energy Phys. 2016 (9) (2016) 1–50.
- [31] L. Collaboration, et al., First observation of the decay $B^0 \rightarrow D^0 \bar{D}^0 K^+ \pi^-$, Phys. Rev. D 102 (5) (2020) 051102, Supplemental Material.
- [32] Dask Development Team, Dask: Library for dynamic task scheduling, 2016, URL <https://dask.org>.
- [33] JLab scientific computing, URL <https://scicomp.jlab.org/scicomp/home>.
- [34] X. Liu, J. Sun, L. Zheng, S. Wang, Y. Liu, T. Wei, Parallelization and optimization of NSGA-II on sunway TaihuLight system, IEEE Trans. Parallel Distrib. Syst. 32 (4) (2020) 975–987.