NVIDIA DGX Infrastructure Solution for Drug Discovery

White Paper

WP-11016-001| July 2022

Document History

WP-11016-001

Version	Date	Authors	Description of Change
1	2022-07-11	Colin Compas, Greg Zynda, John Sanson, Nima Pour Nejatian, Robert Sohigian, Rory Kelleher, Venkatesh Mysore, and Will Vick	Initial Release
2	2022-07-19	Robert Sohigian	Fixed some typos

Abstract

The use of computational methods for multidisciplinary drug discovery is becoming increasingly pervasive. Physics-based simulation and artificial intelligence (AI) are two computational methods that are being leveraged to help advance phenotypic drug discovery (PDD) as well as target-based drug discovery (TDD) programs. These approaches often require significant computing resources to drive the necessary throughput or turnaround time of experimentation and exploration.

Besides supporting current computational drug discovery workflows, extensive in-house compute is being acknowledged as one of the factors essential for enabling breakthroughs. The next set of large algorithmic innovations will further widen the gap in success rate between those who have invested in computational power and those who have not. An unbiased assessment of the scientific landscape suggests that direct access to high performance computing (HPC) and AI infrastructure is a strategic differentiator that enables organizations to use state-of-the-art methods and develop next generation approaches.

Looking at the common applications used across drug discovery pipelines, from sequence to synthesis, GPUs offer the ability to accelerate calculations tens, and in some cases hundreds of times. Both HPC and AI approaches offer the potential to improve time to insight and reduce the overall costs of discovery. The applications of deep learning (DL) in R&D workflows are well understood in the analysis of various biomedical imaging domains as well as natural language processing (NLP). However, DL is now making significant contributions to other areas including genomic variant calling (DeepVariant), protein structure prediction (AlphaFold2, OpenFold), molecule generation and property prediction (GROVER), quantum mechanical energy estimation (OrbNet), as well as models for representing amino acids (ProtTrans) and nucleotides (DNABERT, Enformer).

The ability to support high-throughput in-silico experimentation provides R&D organizations with the tools to glean scientific insights rapidly and cost-effectively. The high-performance applications that are commonly operated at scale include genomics workflows (secondary analysis) and molecular simulation with applications such as GROMACS, Amber, Schrödinger FEP+.

This paper summarizes some of the latest state-of-the-art AI approaches and other high-performance computing applications used across the drug discovery and development process. It also introduces the NVIDIA DGX™ Infrastructure Solution for Drug Discovery (Drug Discovery POD) as the necessary infrastructure to produce desired turnaround times and outcomes for commonly used scientific applications.

Contents

Introduction

The future of drug discovery is computational. Due to advances in simulation techniques and deep learning, more traditional in-vitro methods are being shifted to in-silico approaches. This is reflected in the growing adoption of DL technologies and GPU acceleration of HPC applications leveraged for drug discovery. To address the growing accelerated computing needs of pharmaceutical research and development, NVIDIA has designed the Drug Discovery POD. The Drug Discovery POD is a reference infrastructure design created specifically for pharmaceutical R&D and built around the industry leading NVIDIA DGX A100 system. This white paper outlines leading DL and HPC applications for drug discovery and the impact of the Drug Discovery POD design has on each application's performance at scale. Additionally, this paper provides an overview of the Drug Discovery POD design and how it helps IT organizations simplify deployment, management, and scalability of accelerated compute infrastructure.

State-of-the-Art AI in Drug Discovery Applications

With the advent of new DL approaches based on Transformer architecture, NLP techniques have undergone a revolution in performance and capabilities. Cutting-edge NLP models are becoming the core of modern search engines, voice assistants, chatbots, and more. Modern NLP models can synthesize human-like text and answer questions posed in natural language. There are already examples of pharmaceutical companies using NLP for target identification and prioritization, drug repurposing, and adverse event monitoring.

More recently, these same techniques used in NLP are being used in the interpretation of genes, proteins, small molecules, and other sequentially encoded biochemical data. While these techniques were designed to process language, years of study and data collection allow these models to make new breakthroughs in molecular science.

NLP for Drug Discovery

For pharmaceutical companies, NLP holds the potential to deliver tremendous value for automating text mining, uncovering valuable information hidden among troves of unstructured data. Examples of unstructured data are scientific journal articles, physician notes, medical imaging reports, adverse event reports, and lab notes. In the past, unstructured data was manually analyzed and interpreted. Now, NLP techniques make it possible to create large knowledge graphs with ontology mapping from a massive corpus of textual data. These knowledge graphs can then be used for several downstream tasks, starting with powerful semantic search tools.

BioMegatron: State-of-the-Art Biomedical Language Model

BioMegatron is a language model for biomedical and clinical NLP that was developed at NVIDIA on NVIDIA DGX SuperPOD™. Recent work has demonstrated that large language models dramatically advance NLP applications such as question answering, dialog systems, summarization, and article completion. BioMegatron is the largest biomedical transformer-based language model ever trained, ten times the size of BERT Base, with 345 million, 800 million, and 1.2 billion parameter variants. BioMegatron was trained on 6.1 billion words from PubMed, a repository of abstracts and full text journal articles on biomedical topics. It can be used to extract key concepts and relations from biomedical texts and build knowledge graphs that can drive research and discovery. It can also identify clinical terms in clinical speech and text and map them to a standardized ontology to assist in clinical documentation and research.

Modern NLP models follow a two-step paradigm of pretraining followed by fine-tuning. Pretraining is done on a large corpus of text (for example, PubMed) in an unsupervised manner, producing a scientific language model (for example, BioMegatron). The pretraining process is the most computationally intensive step. This language model is then tweaked for a variety of downstream NLP applications like named entity recognition (NER), relation extraction (RE), and question answering (QA). In the case of domain-specific language models, there is an additional first step of selecting a good vocabulary to train the language model.

To establish the infrastructure requirements for developing a state-of-the-art language model, two different BioMegatron models were pretrained to convergence on different configurations of DGX A100 640 GB systems. Table 1 shows the pretraining benchmarks for the time taken for model convergence on two different model sizes. BioMegatron benchmarks were run using the Megatron-LM v2.2 and PyTorch NVIDIA NGC™ container 21.03 with PubMed full texts and abstracts (total of 6.1 billion words) and a vocabulary size was 30,000 tokens. The number of epochs for convergence for the 345m parameter and 1.2b parameter models were 250 and 120, respectively. All training runs were completed using NVIDIA Megatron-LM framework.

Table 1. BioMegatron pretraining training time.

Model Size	No. of DGX A100 Nodes	Approximate Total Training Time to Convergence (days)
345m (L:24, H:1024, A:16)[1]	2	33
	8	8.5
	20	3
	40	1.5
1.2b (L:24, H:2048, A:32)	2	45
	8	11.5
	20	4.5
	40	2.5
1. L: The number of transformer Layers, H: Hidden unit size, A: the number of Attention heads		

Transformers For Other Biomolecules

Self-supervised approaches, such as in transformer models, have shown the ability to provide informative representations of complex biomolecules including small molecules, amino acids, peptides, nucleotides and more. This is particularly useful for sparse data scenarios where experimental datasets are limited. These embedding models can aid in a variety of downstream biomolecule AI tasks—from property and binding prediction to de novo generation. Below, we highlight the recent literature in applying these self-supervised Transformer models for small molecules, amino acids, and nucleotides, based on recent literature.

Small Molecule Transformers

Transformers have shown the ability to embed the dense and continuous representation of chemical space, which adheres to the rules of chemistry and captures the underlying distribution of molecules in the database. This confers the correct inductive bias needed for one-shot learning and other problems where data is limited, or labels are not available. Downstream applications of these models include property prediction, molecule generation, and more.

A representative application for small molecule transformers is Graph Representation from self-supervised message passing transformer (GROVER). Pretraining of a 100M parameter GROVER model took place on 250 V100 GPUs for ~4 days. Similarly, a 120M parameter GROVER model can be trained on DGX A100 system with the runtimes shown in Table 2.

Table 2. Training throughput of GROVER

Model Architecture	No. of DGX Systems	Time per Epoch (minutes)
GROVER 120M	1	62
	2	31.5
	10	6.76

Amino Acid Transformers

Biological data sets such as amino acid sequences are ideal data sets for structure prediction, protein-protein interactions, protein small molecule interactions, enzyme design, and antibody design. A representative workload for amino acid transformers is the ProtTrans networks. The ProTXL-BFD was trained on 5,616 V100 GPUs for five days.

Quantum Mechanical Energy Estimation

Physics-informed neural networks (PINNs), or DL algorithms that combine mathematical models and data, are giving rise to new capabilities in biological and organic chemistry. OrbNet–Denali is one such model published by Entos/Caltech, which predicts DFT energies and forces to drive molecular simulation and to predict molecular physical properties. This DL approach to DFT offers a speedup of three orders of magnitude compared to other modern DFT methods.

OrbNet Denali was trained at Oak Ridge Leadership Computing Facility Summit Supercomputer with 96 V100 (32G) GPUs for three days. Additional details of the training parameters can be found in the reference publication. Leveraging the high-performance environment of Drug Discovery POD, there should be significant performance improvements for OrbNet Denali.

Deep Learning Powered Image Analysis

Image analysis provides key insights in multiple stages of a drug discovery pipeline. From determining the structure of biological macromolecules with cryo-EM, to high-content screening (HCS) for compound testing, to monitoring disease progression after treatment in diagnostic imaging. All of these imaging modalities are benefiting from advances in DL to analyze the large volumes of data being produced. A common task in image analysis is classification. In diagnostic imaging, this could be identifying the presence or absence of a tumor. In HCS, this could be identifying cellular response to a compound. EfficientNet is convolutional neural network architecture that achieves state-of-the-art performance on different classification tasks. To measure the performance of EfficientNet v2-S, the implementation provided in NVIDIA Deep Learning Examples was used. EfficientNet v2-S was implemented in TensorFlow 2 using the TensorFlow 21.09-py3 container from ngc.nvidia.com and trained on the ImageNet dataset. Full details of the implementation can be found in the GitHub repository.

Table 3 shows the result of this benchmarking for various DGX configurations up to ten DGX A100 systems.

Table 3. Time to train EfficientNet v2-S

Model Architecture	No. of DGX Systems	Total Training Time to Convergence (hours)
EfficientNet v2-S	1	14
	2	7
	10	1.8

Pathology also produces high-resolution images of cells and tissues on the order of gigabytes per file. Because these images are so large, they are usually processed in "patches," or fixed-sized subselections. NVIDIA has worked to accelerate this type of workflow from a standard ResNet-18 model to one with accelerated image loading, image transformations, and automatic mixed-precision (AMP) to greatly accelerate this type of work on DGX A100 system. The speedups shown in Table 4 are achieved through cuCIM to accelerate image loading and transformations, and AMP that is built in to PyTorch.

Table 4. Comparison of GPU-optimized throughput of Histopathology workflow

Model Architecture	No. of DGX Systems	Unoptimized (patches/second)	GPU-Optimized (patches/second)
ResNet-18	1	307	2178
	2	608	3603
	10	1475	10474

High-Throughput Workflows in Drug Discovery

The efficiency and productivity of modern drug discovery can be significantly improved using in-silico methods. Drug programs that start with a genomic understanding of disease are twice as likely to succeed in the clinic. As such, pharma companies have invested in sequencing efforts and downstream interpretation of that data to determine causal biology. In addition, the ability to rapidly search chemical space and conduct computational predictions of molecule characteristics or binding energies at near-experimental accuracies is driving the cost and speed of drug discovery to new levels.

Genomics: Secondary Analysis

GPU-accelerated NVIDIA Clara™ Parabricks generate results are 30–60X faster than industry-standard CPU-based workflows for DNA and RNA analysis. Clara Parabricks accelerates germline analysis of both GATK4.1 and Google DeepVariant along with a suite of somatic callers including Strelka2, LoFreq, SomaticSniper, and Mutect2, all reimplemented to run on a suite of NVIDIA GPU platforms, delivering the same results as the native CPU instances but in minutes instead of hours or days.

When calling genomic variants from raw FASTQ sequence files, a single DGX A100 node can process the whole genome HG002 reference sample, with 30x coverage, using the GATK HaplotypeCaller in 22 minutes and Google AI DeepVariant in 30 minutes. As analysis demands increase, Table 5 shows how the performance of Clara Parabricks scales over DGX A100 systems for samples with the same sequencing depth.

Table 5. Clara Parabricks GATK HaplotypeCaller runtime

Coverage Depth	No. of DGX Systems	Samples per year
Germline 30x Coverage on Illumina short read for whole genome	2	48,000
	4	96,000
	10	240,000

Molecular Simulation

The demands of simulation in target-based drug discovery have increased significantly, due to two recent breakthroughs. The first is the advancement of protein structure prediction tools such as AlphaFold2, which have provided a wealth of target proteins. The second is the creation of ultra-large libraries combined with generative methods to create novel molecules outside of databases that are pushing the computational bottleneck to in-silico simulations.

Molecular simulation has been used for molecule design for decades. The majority of popular molecular dynamics simulations packages are accelerated on GPUs including AMBER, GROMACS, VASP, free energy perturbation (FEP+), and OpenMM. Determining binding affinities of molecules to target proteins, as well as off-target proteins, is an important and challenging step in determining the best molecules to synthesize. Current techniques such as absolute binding free energy calculations combined with the massive computational power of GPUs are now showing experimental accuracy which is promising for research and clinical use.

The following section shows benchmark results for commonly used MD applications on DGX A100 hardware configurations.

GROMACS

GROMACS is one of the most widely used codes in chemistry, used primarily for dynamical simulations of biomolecules. Several advanced techniques for free-energy calculations are supported. In version 5, it reaches new performance heights, through several new and enhanced algorithms. These work on every level—SIMD registration inside cores, multithreading, heterogeneous CPU–GPU acceleration, state-of-the-art 3D, and ensemble-level parallelization through built-in replica exchanges and separate Copernicus frameworks.

Table 6 details the performance of GROMACS on two different system sizes for 24k and 96k atoms across different DGX A100 640 GB configurations up to one SuperPOD where 300,000 ns/day can be simulated. These simulation numbers leverage the latest development optimizations with MIG and MPS that are outlined in, Maximizing GROMACS Throughput with Multiple Simulations per GPU Using MPS and MIG.

Table 6. GROMACS performance on DGX A100 systems

System Size	No. of DGX Systems	(ns / day)
24k atoms	1	15,000
	4	60,000
	10	150,000
96k atoms	1	3,700
	4	14,800
	10	37,000

AMBER

Assisted Model Building with Energy Refinement (AMBER) is a suite of biomolecular simulation programs that was created in the late 1970s and still has an active development community.

The ability to simulate for several microseconds of chemical time, rather than several nanoseconds, opens the possibility of observing several phenomena critical to structure-based drug discovery. Long MD simulations have been a reliable way of studying binding site flexibility, from side-chain torsional flips to the coordinated motion of a set of residues. These often reveal side-pockets that can be exploited for potency, novelty, and specificity. Similarly, loop regions and long linker regions often exhibit conformational changes in the whole protein which can reveal allosteric binding sites.

Numerous studies have also shown that microsecond simulations can be used for de novo ligand and lipid binding for previously uncharacterized sites and unbiased induced fit of known ligands to apo structures. Being able to simulate large systems becomes critical for studying the structure and function of multidomain proteins, protein-protein complexes and membrane proteins, which together comprise a large fraction of hot pharmaceutical targets. Thus, the ability to run longer MD simulations of larger systems can confer a key competitive edge to any structure-based drug discovery enterprise.

Table 7 shows the performance of AMBER on a real-world workload made up of with a mix of 20% Cellulose (408,609 atoms), 40% FactorIX (90,906 atoms) and 40% JAC (23,558 atoms) run across different DGX A100 640 GB configurations.

Table 7. AMBER performance on DGX A100 systems

No. of DGX Systems	(ns/day)
2	6,800
4	13,600
10	34,000

Schrödinger FEP+

Computing relative binding FEP+ provides significant value to industrial drug discovery efforts by reducing the time spent in lead optimization. Lead optimization is considered the most expensive phase of drug discovery since it is the process by which a drug candidate is iteratively synthesized and characterized after a lead compound is identified. FEP calculations use molecular simulation to predict ligand-protein binding affinities during lead optimization, and GPU acceleration has helped FEP+ reduce the duration of this phase even further.

Starting with SLURM v21.08, the Schrödinger job server can schedule tasks on MIG instances. The FEP+ workflow alternates between CPU and GPU computation, and the increased task density allows for an average of 3.5x increase in simulation throughput over tasks scheduled on whole A100 GPUs. BACE-1 (1-edge) simulations have a maximum throughput of 128.8 perturbations/day per DGX when each task is allocated two 1g.10gb MIG instances. Larger, 6-edge simulations run at 35 perturbations/day per DGX A100 system when each task is allocated four 1g.10gb MIG instances.

The maximum throughput of BACE-1 and 6-edge simulations using FEP+ when scheduled on 1g.10gb MIG instances for various numbers of DGX A100 systems is in Table 8.

Table 8. FEP+ throughput on DGX A100 systems

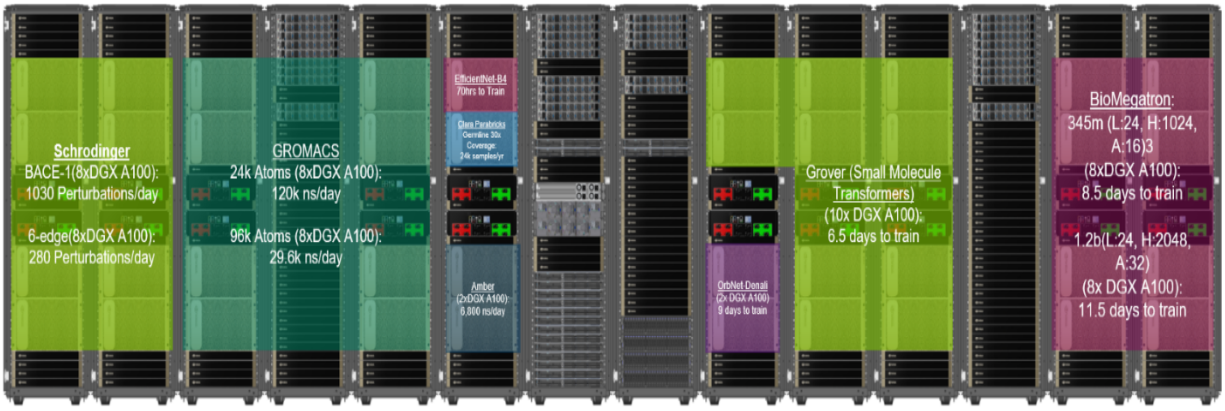
Benchmark	No. of DGX Systems	(perturbations/day)
BACE-1	2	257
	4	515
	10	1288
6-edge	2	70
	4	140
	10	350

Next Generation Drug Discovery Requires Leadership-Class Computing Infrastructure

Designing and building scaled computing infrastructure for AI requires an understanding of the computing goals of AI researchers in order to build fast, capable, and cost-efficient systems. Developing infrastructure requirements can often be difficult because the needs of research are often an ever-moving target and AI models, and due to their proprietary nature, often cannot be shared with vendors. Additionally, crafting robust benchmarks that represent the overall needs of an organization is a time-consuming process. This dilemma creates the requirement for organizations to leverage a standardized approach to building and scaling AI infrastructure. The Drug Discovery POD is designed to help pharmaceutical research organizations overcome the difficulties in designing, deploying, operating, and maintaining a best-in-class operating environment for the most intensive AI and HPC workloads. As the leader in accelerated computing and data science, NVIDIA has leveraged experience from deploying solutions with the leading pharmaceutical and research organizations to build the Drug Discovery POD. This solution takes the focus off IT and allows researchers to focus on their most important health initiatives.

The Drug Discovery POD is built upon the NVIDIA DGX Stack with Domino Data Labs, with a sample configuration shown in Figure 1.

Figure 1. Sample Drug Discovery POD configuration



The drug discovery applications highlighted in this document are tested and benchmarked on this validated infrastructure platform by leading HLS engineers at NVIDIA. Built around the world class DGX A100 system and following the NVIDIA DGX POD architecture, IT departments follow a prescriptive process to scale your AI environment to meet the most intensive performance needs in the world. To make deployment and management seamless and efficient, the Drug Discovery POD comes with NVIDIA Bright Cluster Manager platform. Bright Cluster Manager allows IT administrators to streamline deployment, management, and monitoring removing risk and delays involved in building out enterprise GPU enabled clusters.

As with any DGX investment, the Drug Discovery POD provides users access and support to a plethora of NVIDIA developed optimized applications and toolkits with NVIDIA AI Enterprise and NGC offering a curated set of NVIDIA applications and integrations that simplify building, customizing, and integration of GPU-optimized software into workflows, accelerating the time to productivity for users. Solutions such as GROMACS, NVIDIA Clara GPU-accelerated tools, Schrödinger, EfficientNetv2, and NVIDIA BioMegatron are accelerated with applications in the NGC catalog. For data scientists and researchers, the technology stack includes Domino Data Labs Enterprise as an optional solution for MLOPs. Domino Data Labs provides users with a MLOPs, data science and experiment management environment that helps tackle the most challenging AI projects. With the Drug Discovery POD, IT departments, data scientist and researchers can build an AI Center of Excellence with a NVIDIA validated technology stack to drive all GPU enabled workloads for drug discovery.

With the NVIDIA DGX Solution for Drug Discovery, organizations can expect:

- A fully validated operating environment for infrastructure management, data science, and research.
- NVIDIA engineered solution architected for streamlined scalability and predictable performance.
- Powerful scale-up nodes, a large memory footprint, and fast connections between the GPUs for computing to support the variety of DL models and HPC applications.
- A low-latency, high-bandwidth, network interconnect designed with the capacity and topology to minimize bottlenecks.
- Support and access from leading NVIDIA engineers to drive initiatives.
- DGX workshops for startup and operations for IT Administrators.

Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation (“NVIDIA”) makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer (“Terms of Sale”). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA DGX, NVIDIA DGX SuperPOD, NVIDIA Clara, and NVIDIA NGC are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com