A multi-objective perspective for the Once-for-All neural architecture search framework

MSc Defense

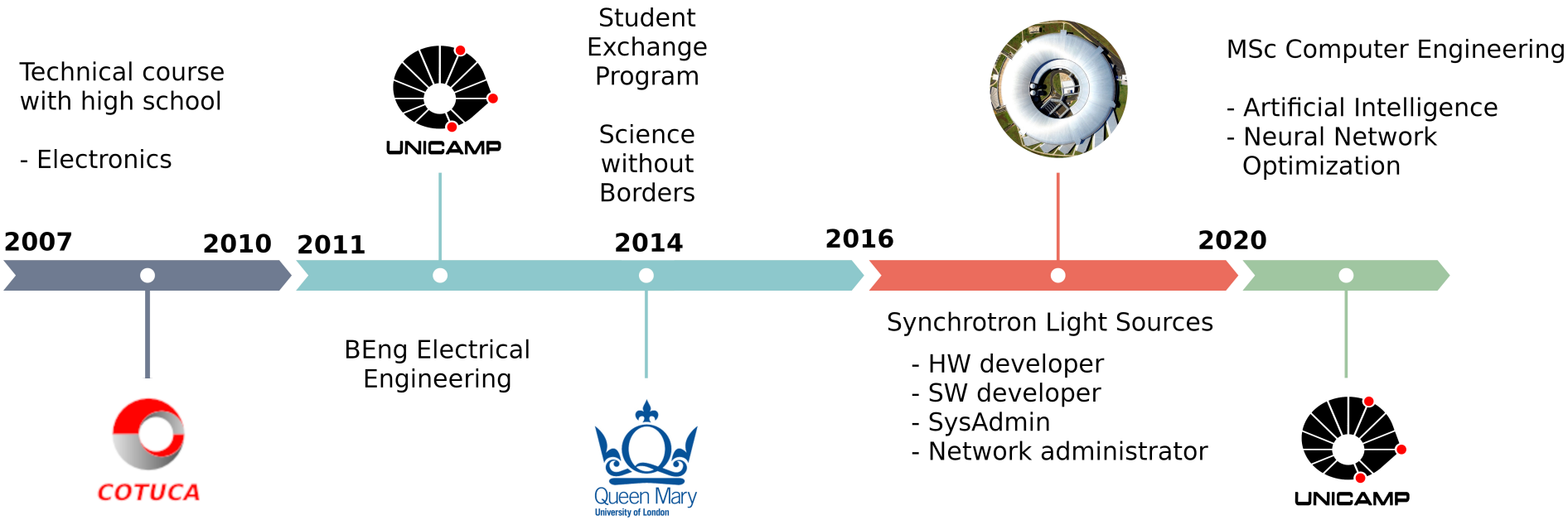
Rafael Claro Ito

28/02/2023

Outline

Agenda

whoami?

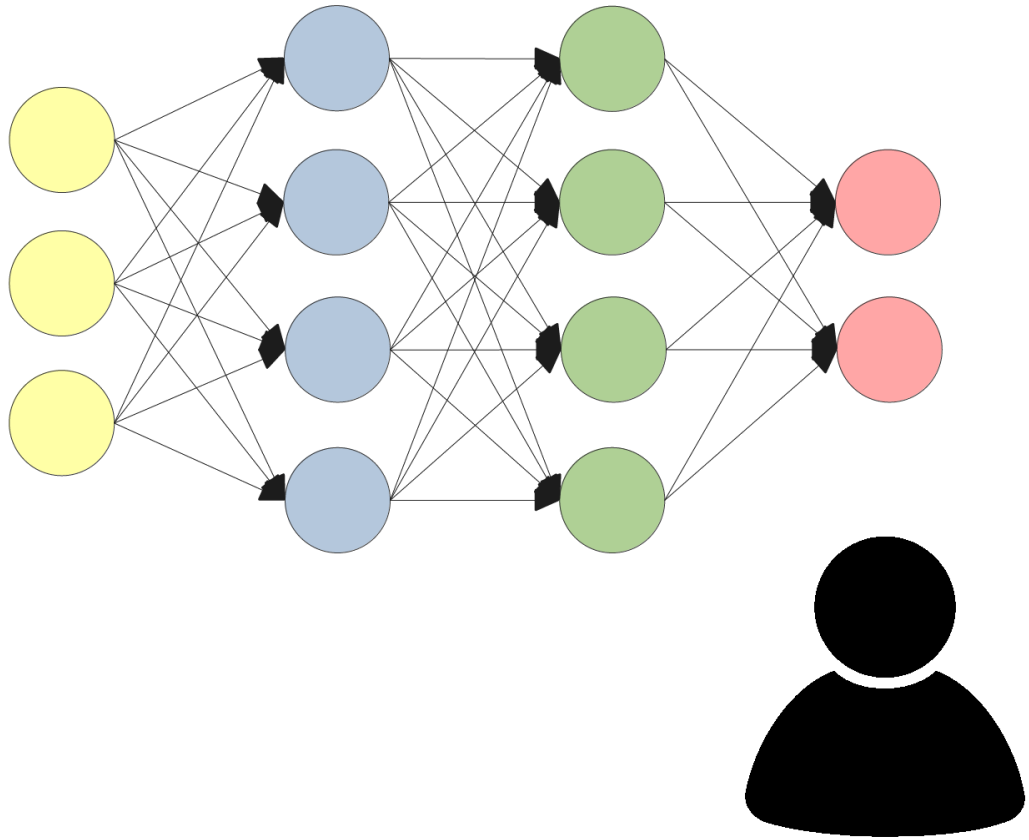

Technical course with high school
- Electronics
2007
2010
COTUCA
UNICAMP
2011
BEng Electrical Engineering
Student Exchange Program
Science without Borders
2014
Queen Mary University of London
2016
Synchrotron Light Sources
- HW developer
- SW developer
- SysAdmin
- Network administrator
MSc Computer Engineering
- Artificial Intelligence
- Neural Network Optimization
2020
UNICAMP

Acknowledgements

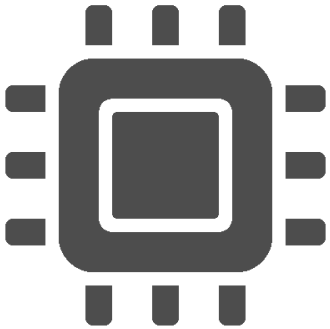
Part of the results presented in this work were obtained through the project "Hub of Artificial Intelligence in Health and Wellbeing – Viva Bem", funded by Samsung Eletrônica da Amazônia Ltda.

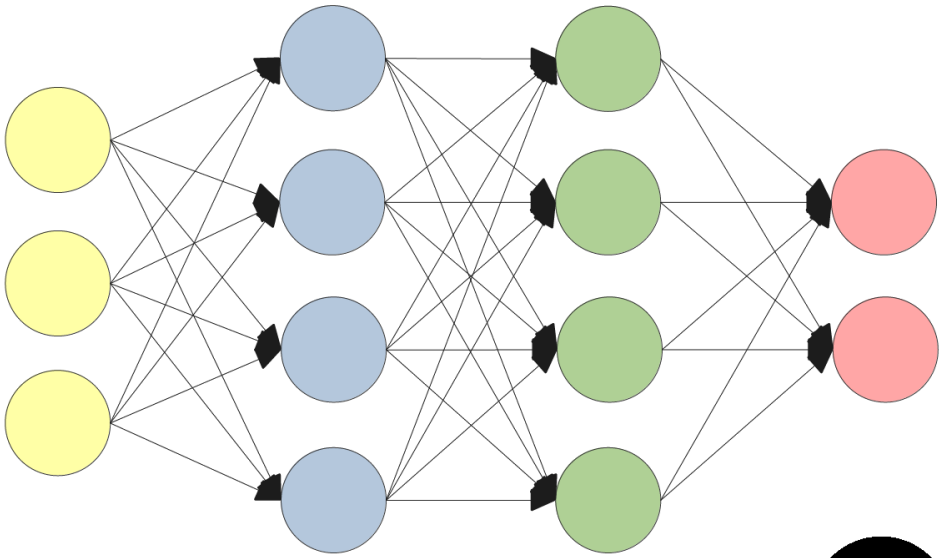
Hub of Artificial Intelligence for Health and Well-being.

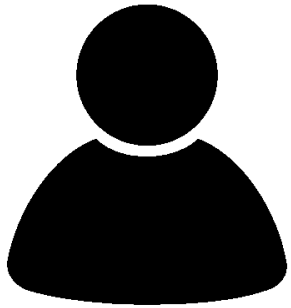
One deployment scenario



One deployment scenario



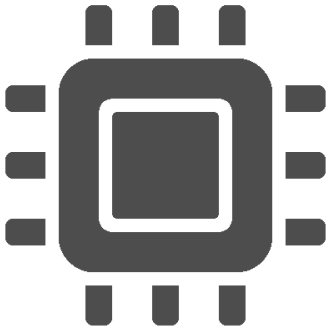


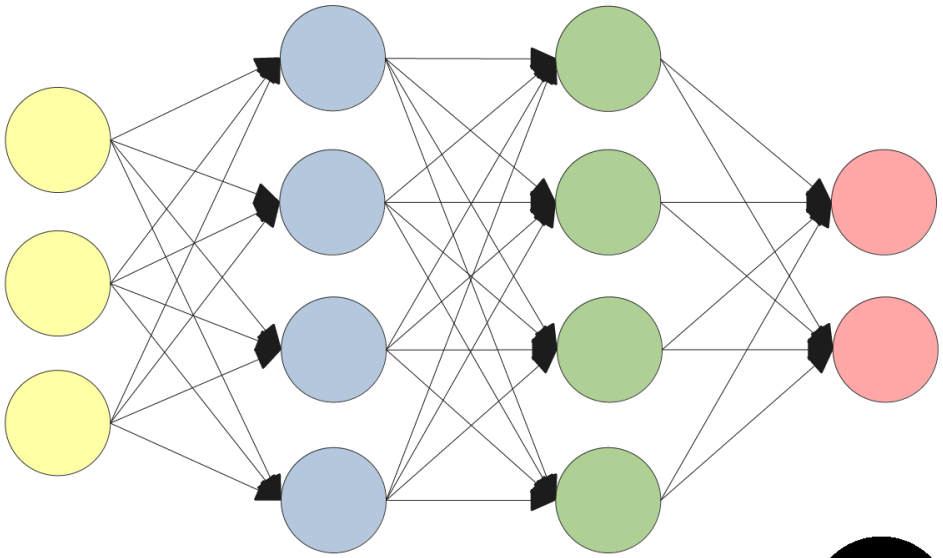


Hardware restriction

Metric	Example
memory	12 GB
latency	50 ms
FLOPS	1 G
MACs	600 M

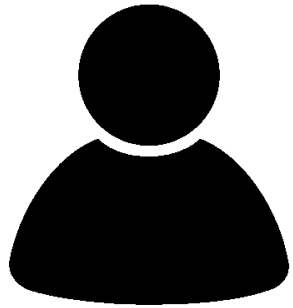
One deployment scenario



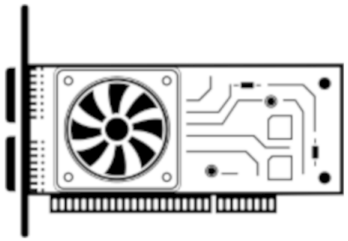


Hardware restriction

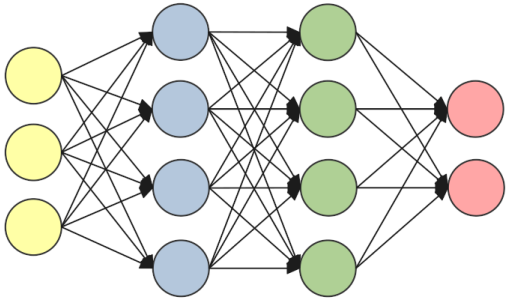
Metric	Example
memory	12 GB
latency	50 ms
FLOPS	1 G
MACs	600 M



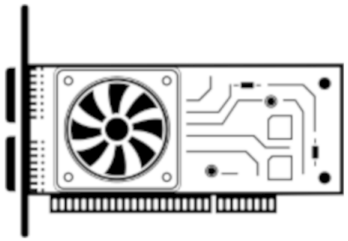
Many deployment scenario



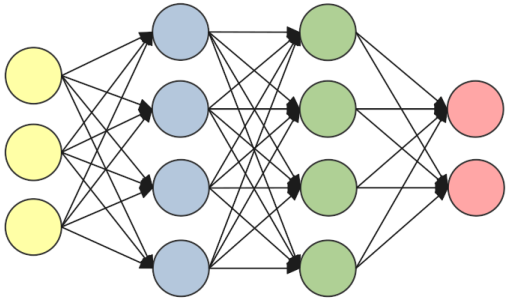
GPU

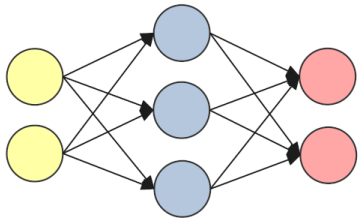


Many deployment scenario

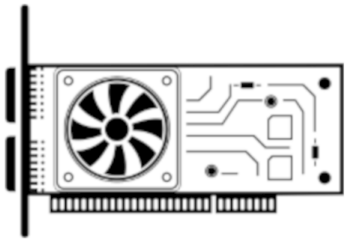


GPU

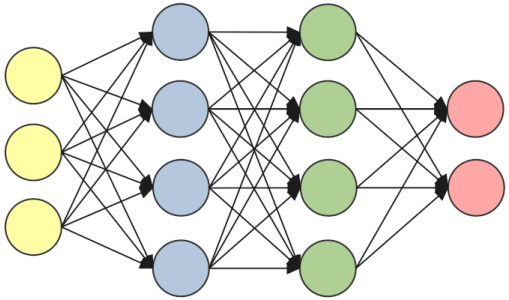


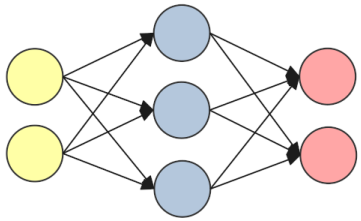


Many deployment scenario

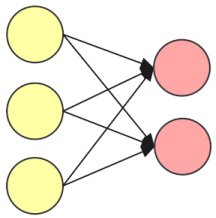


GPU

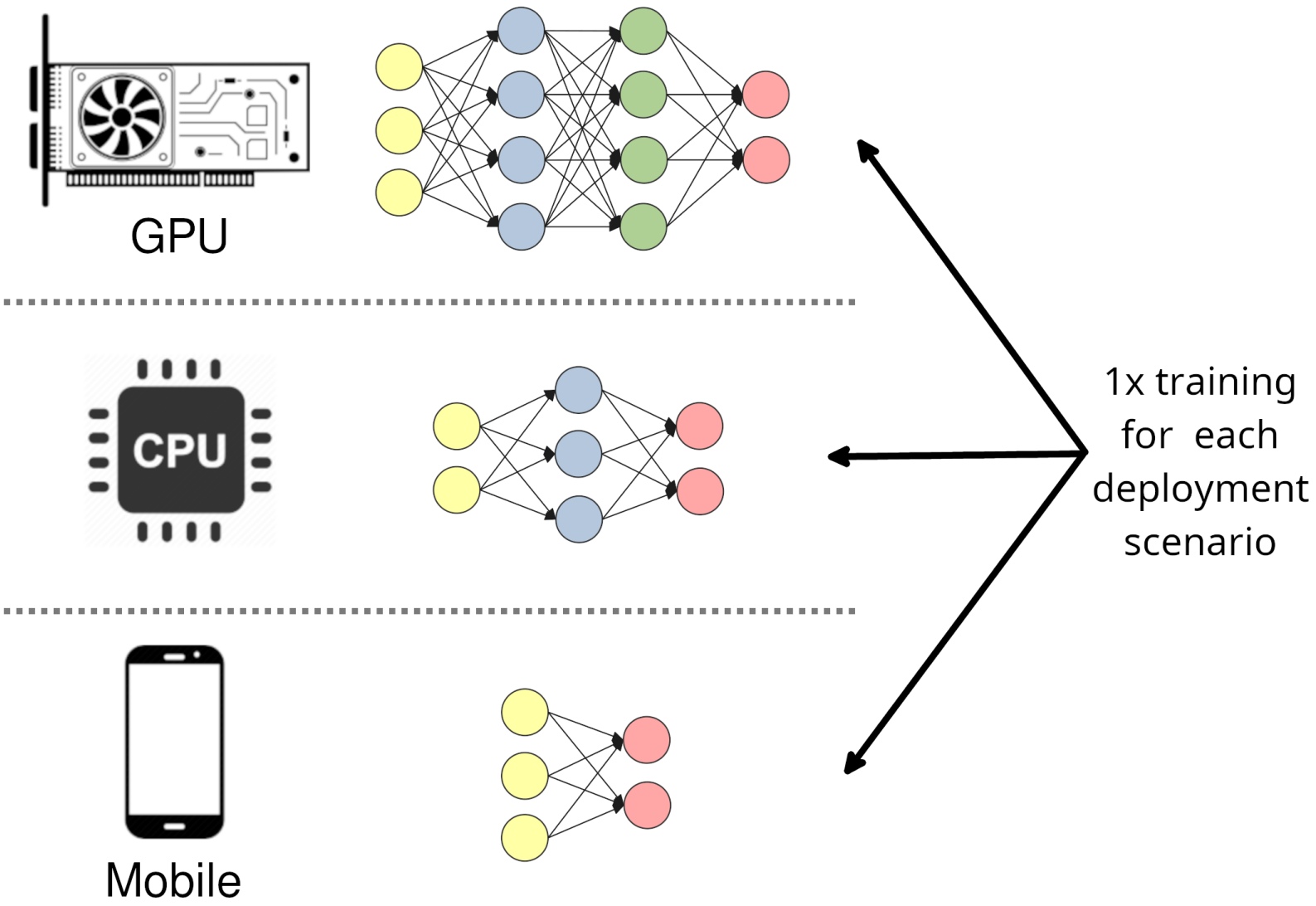




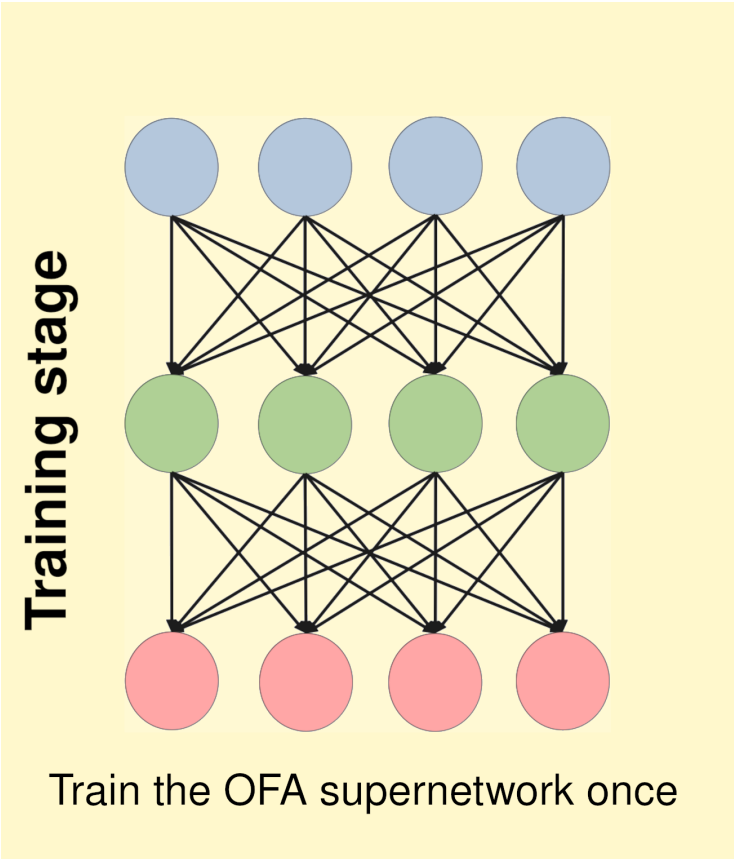
Mobile



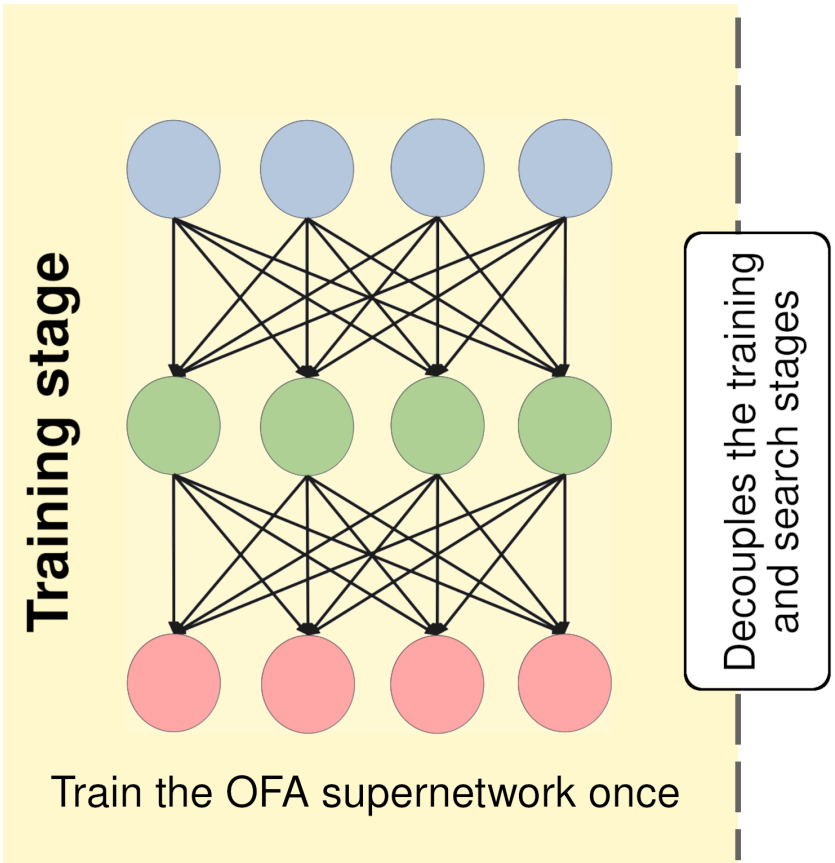
Many deployment scenario



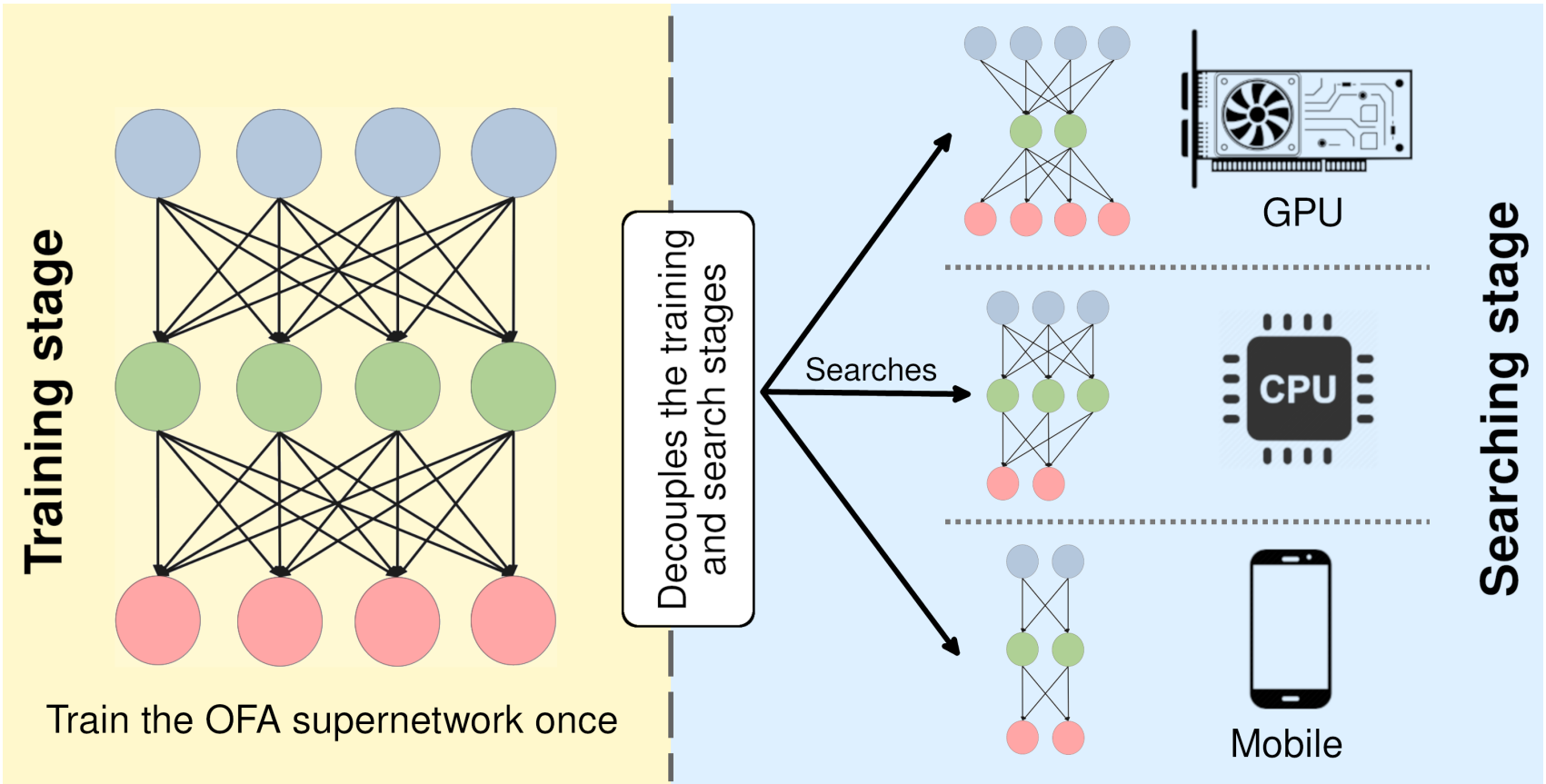
Once-for-All framework



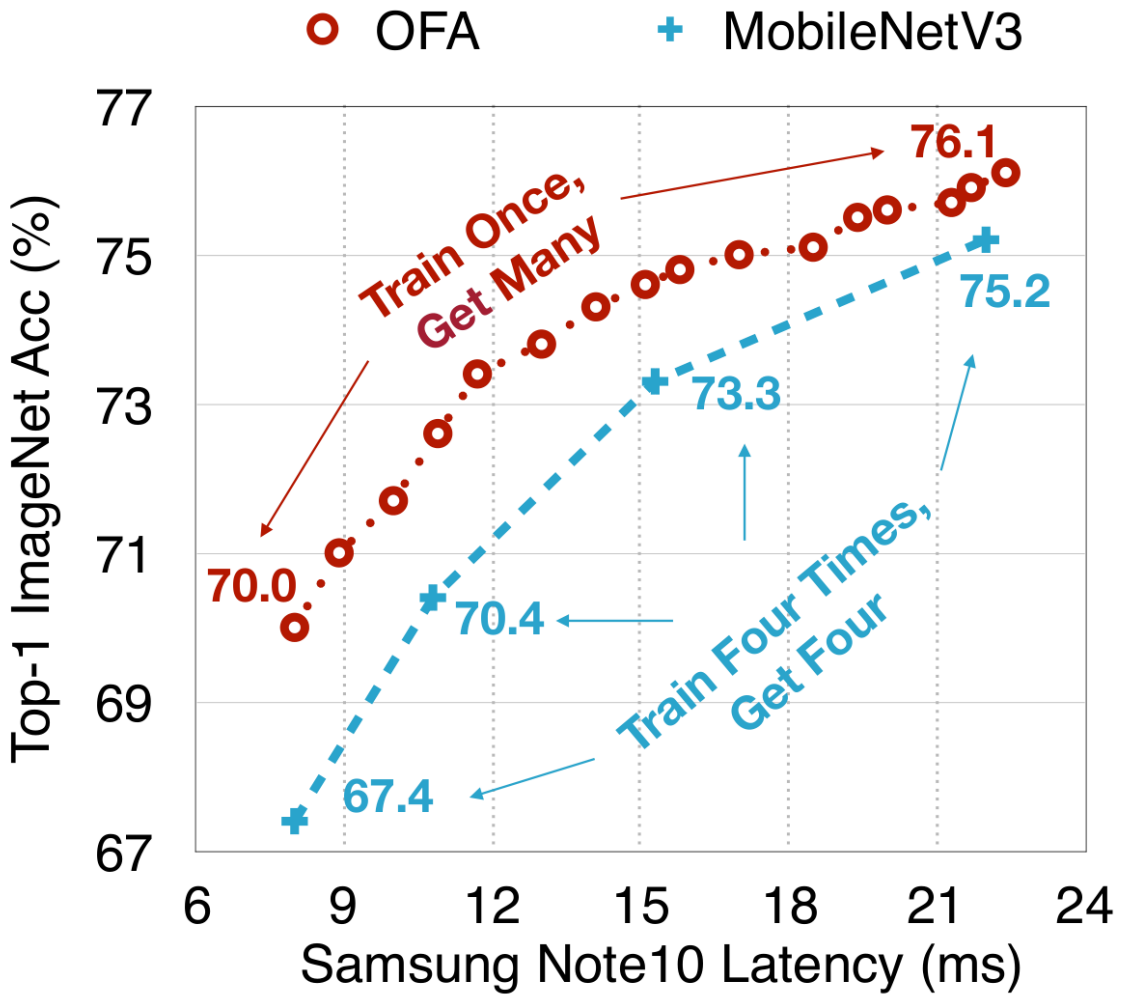
Once-for-All framework



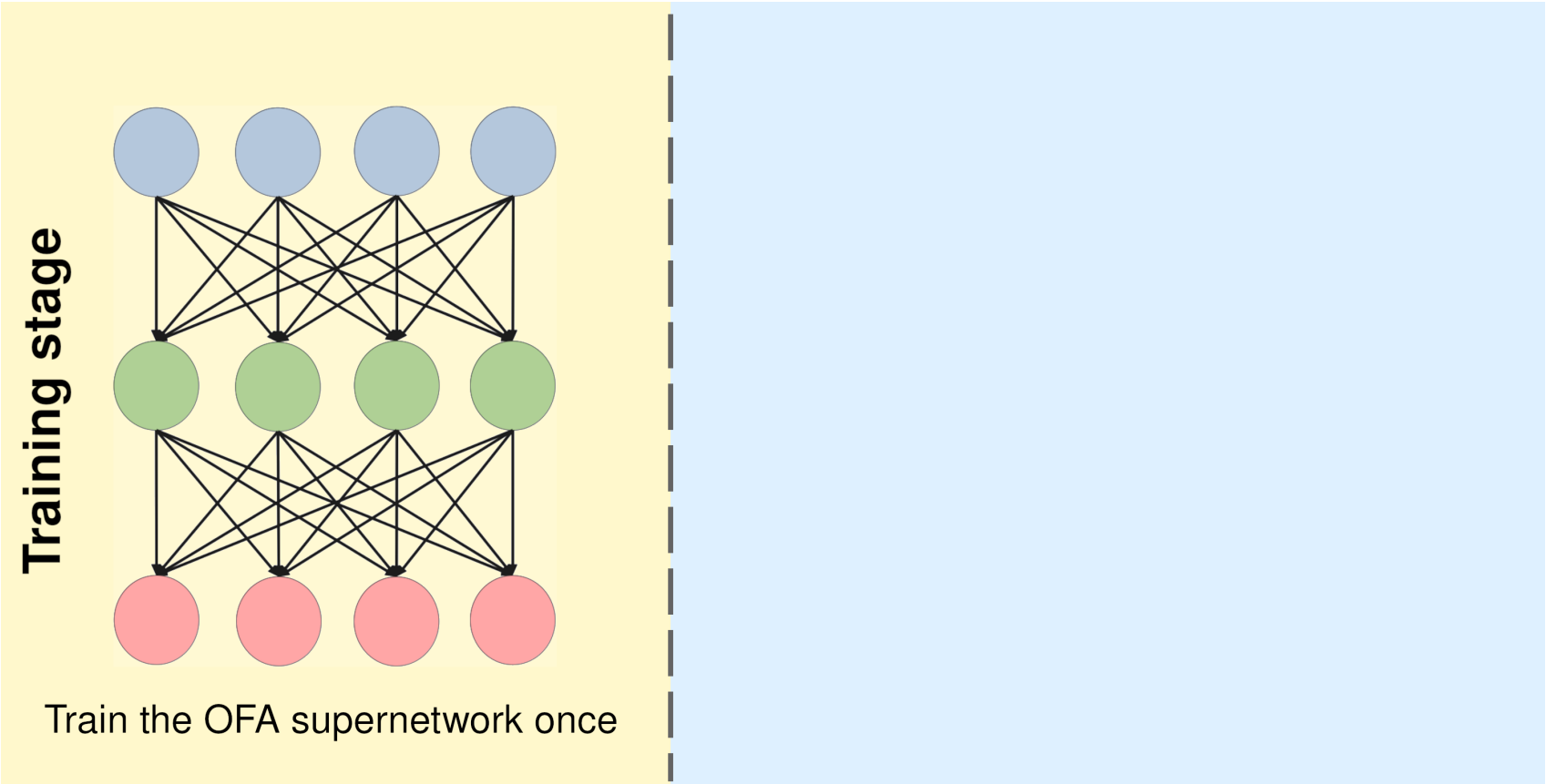
Once-for-All framework



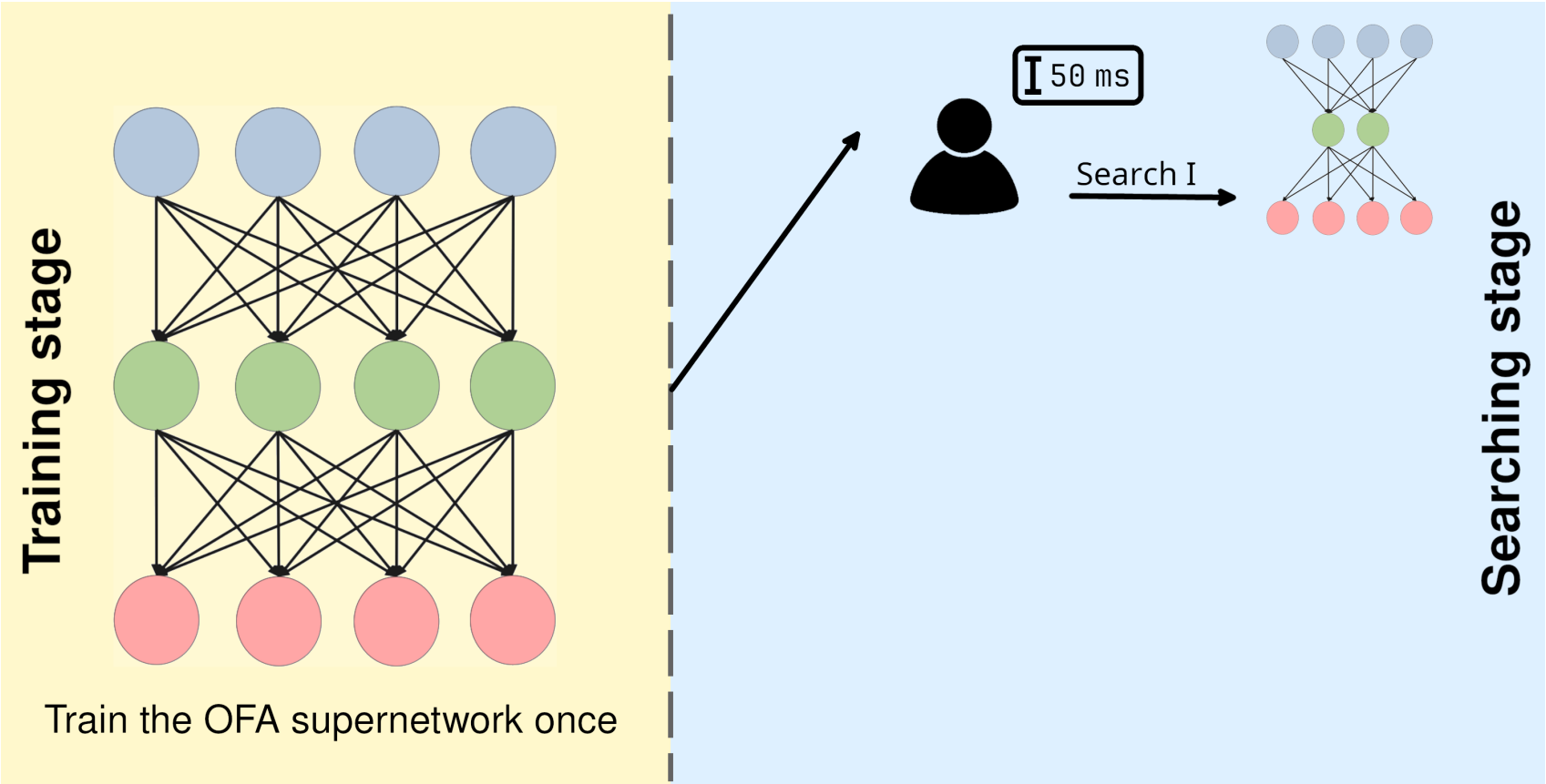
Once-for-All framework



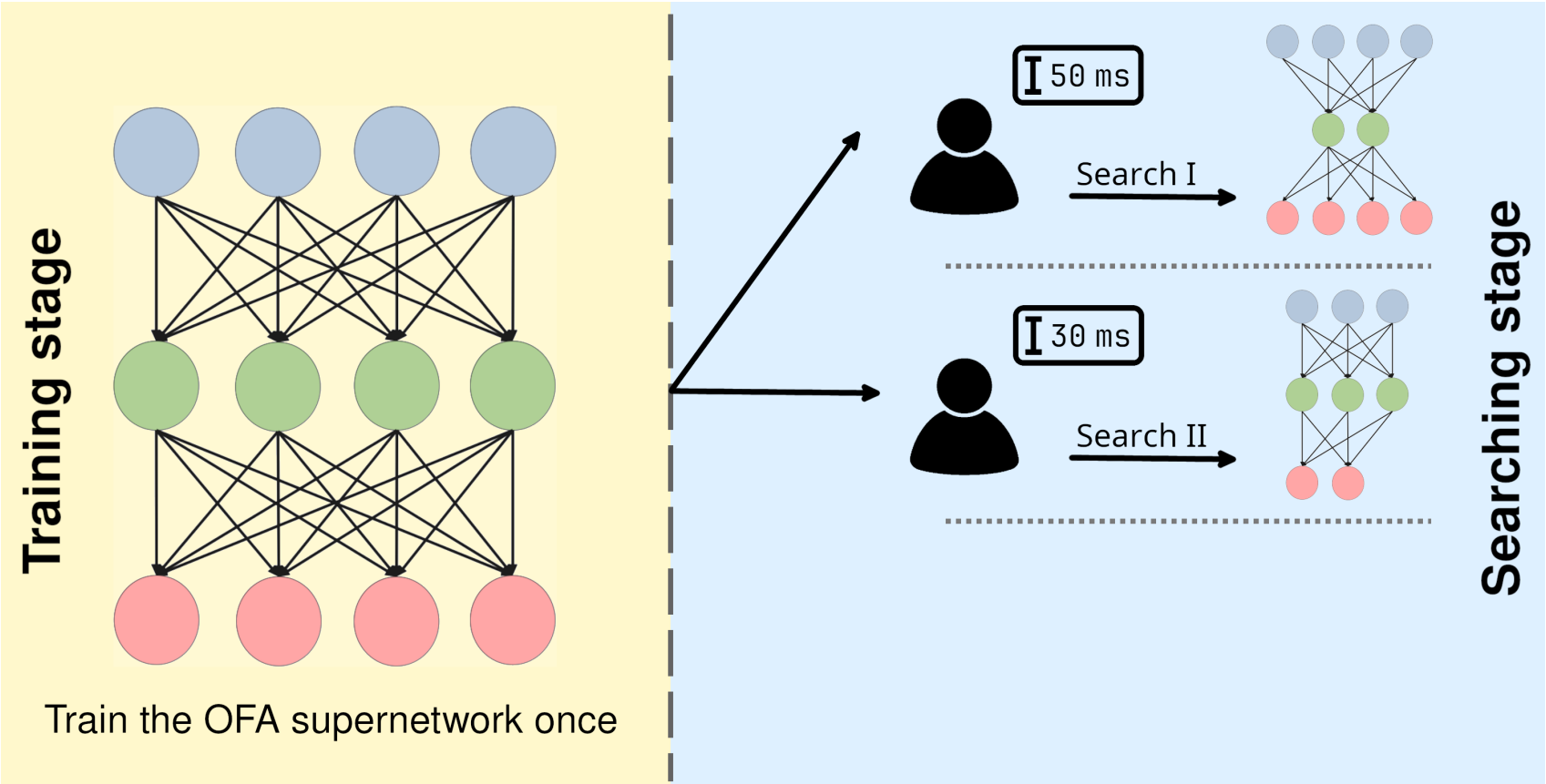
Once-for-All framework



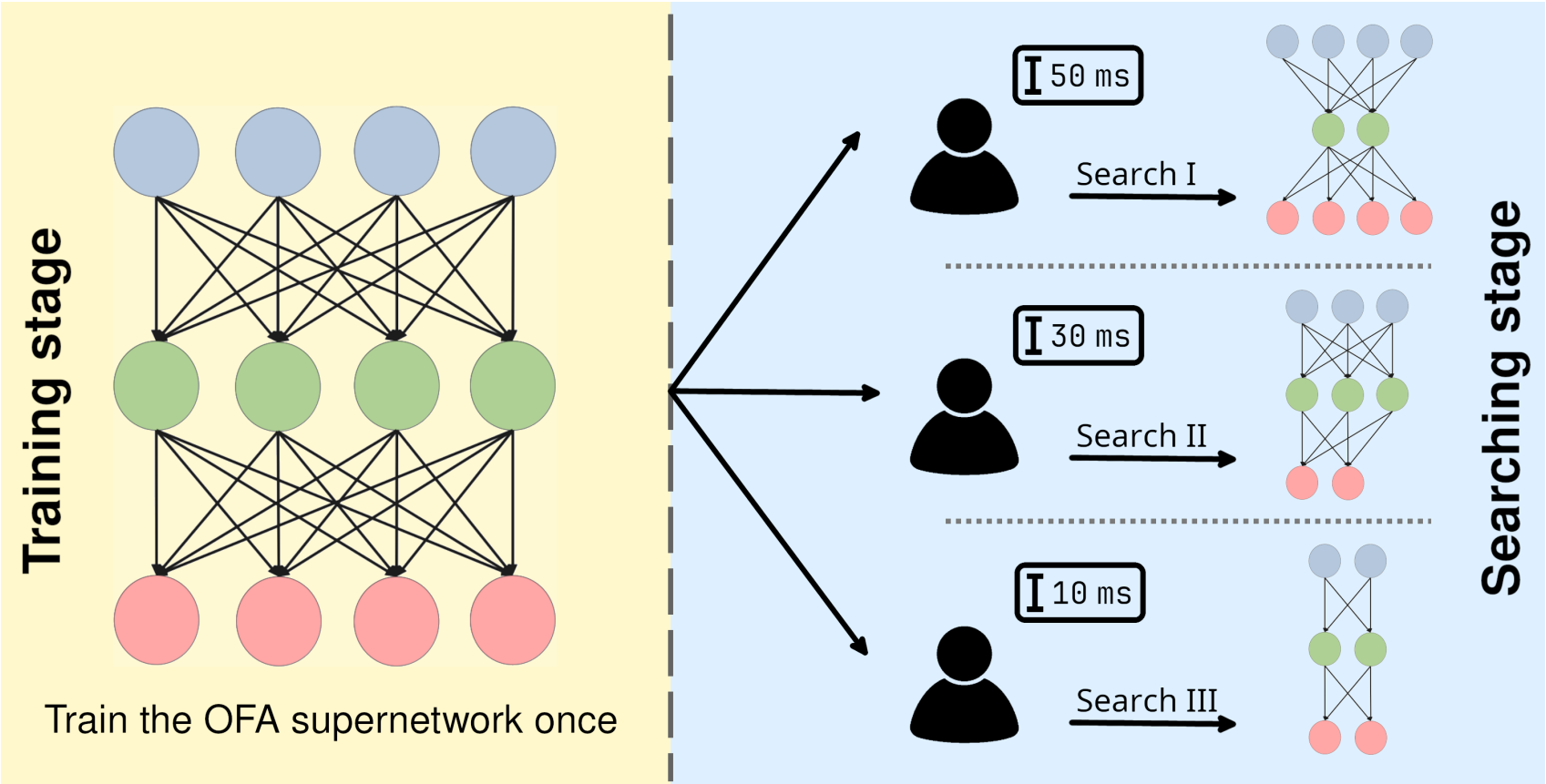
Once-for-All framework



Once-for-All framework

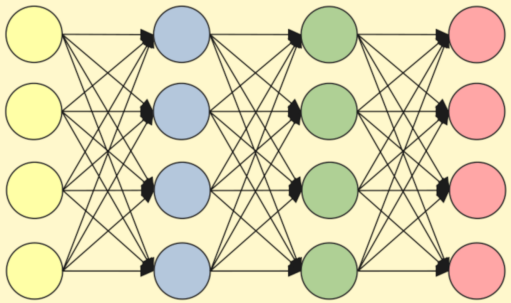


Once-for-All framework



OFA² framework

Once-for-All

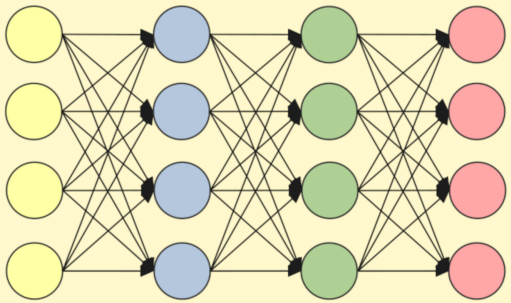


Train the OFA supernetwork once

Training stage

OFA² framework

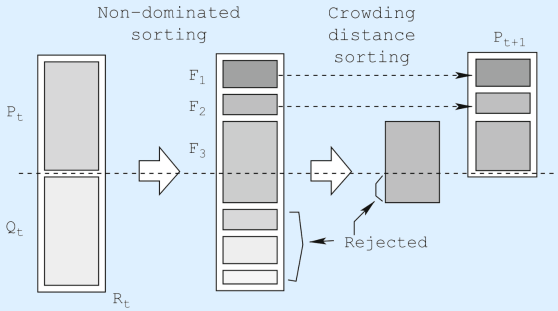
Once-for-All



Train the OFA supernetwork once

Training stage

EMOA

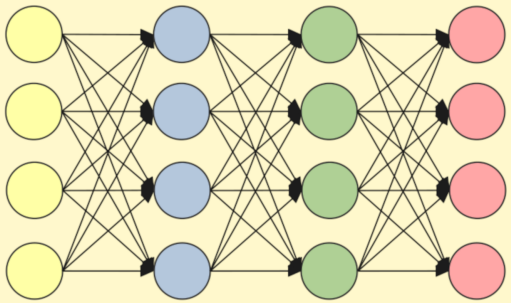


Search once with a MOO formulation

Search stage

OFA² framework

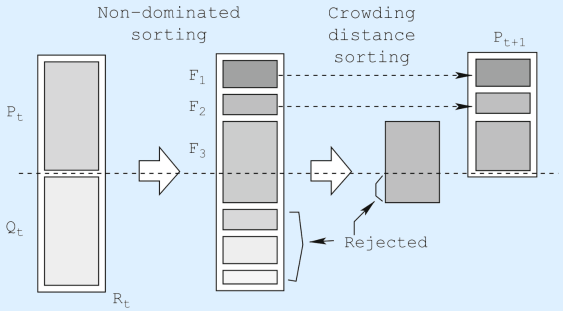
Once-for-All



Train the OFA supernetwork once

Training stage

EMOA

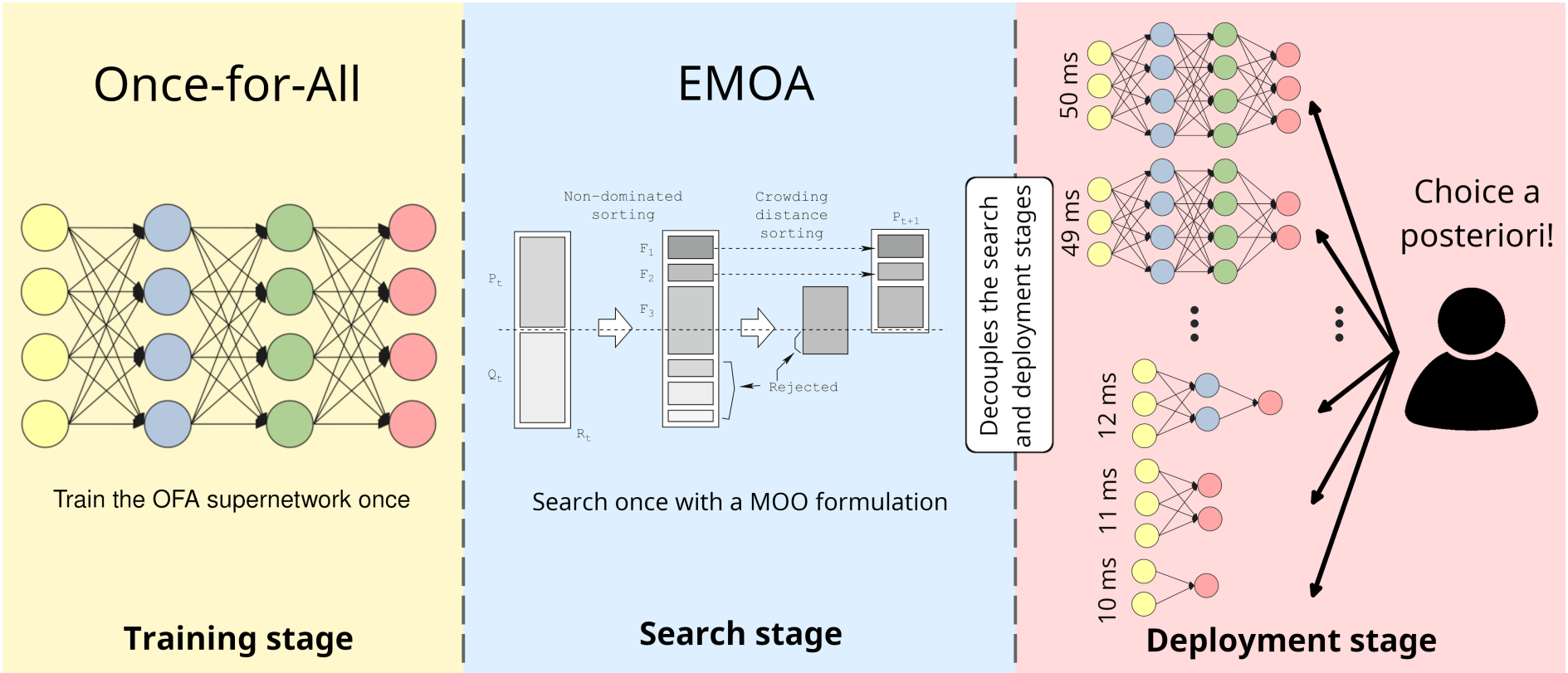


Search once with a MOO formulation

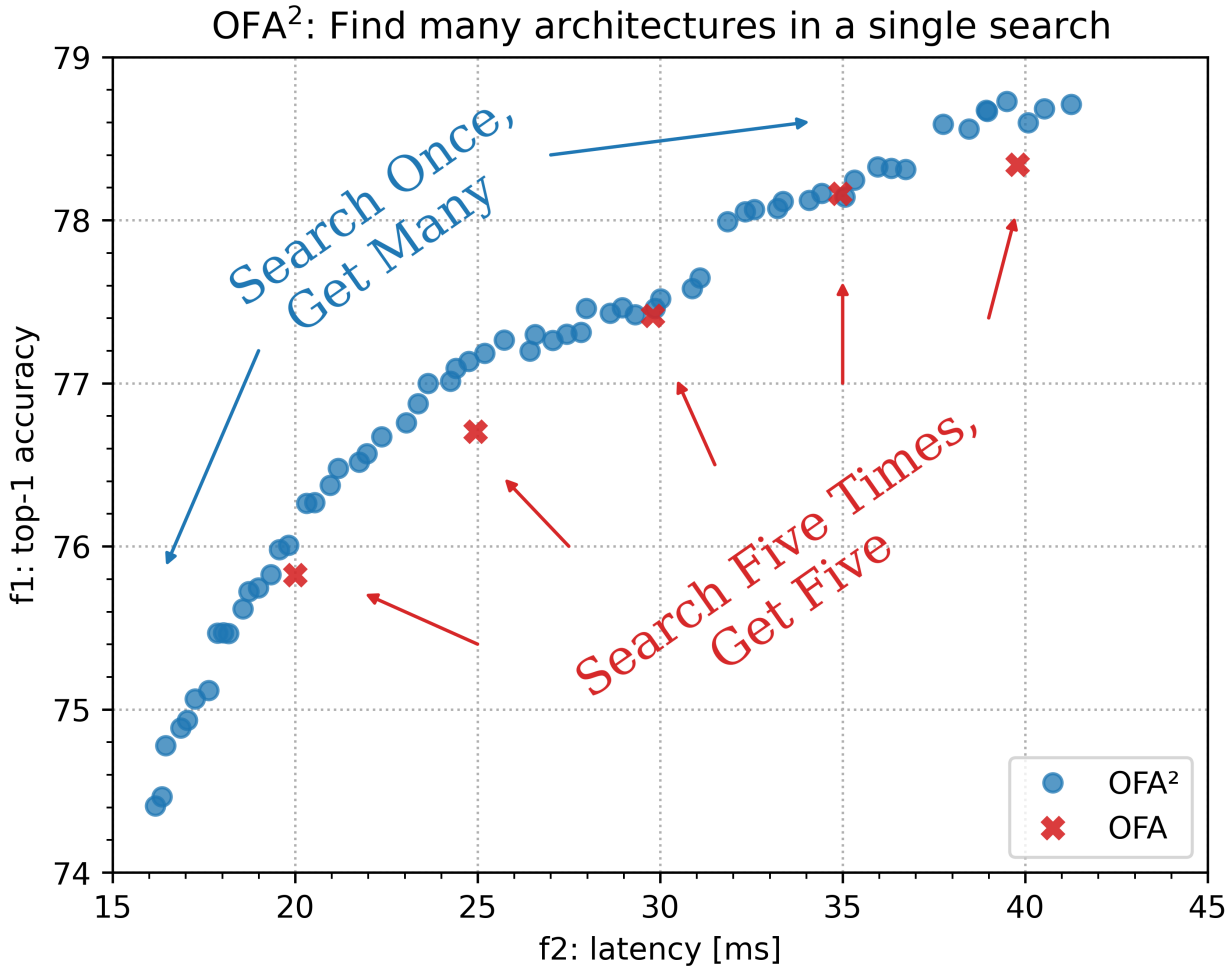
Search stage

Decouples the search and deployment stages

OFA² framework

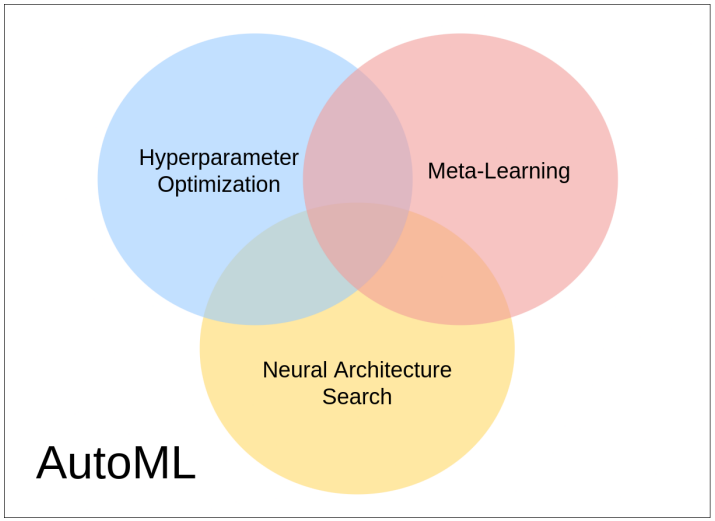


OFA² framework



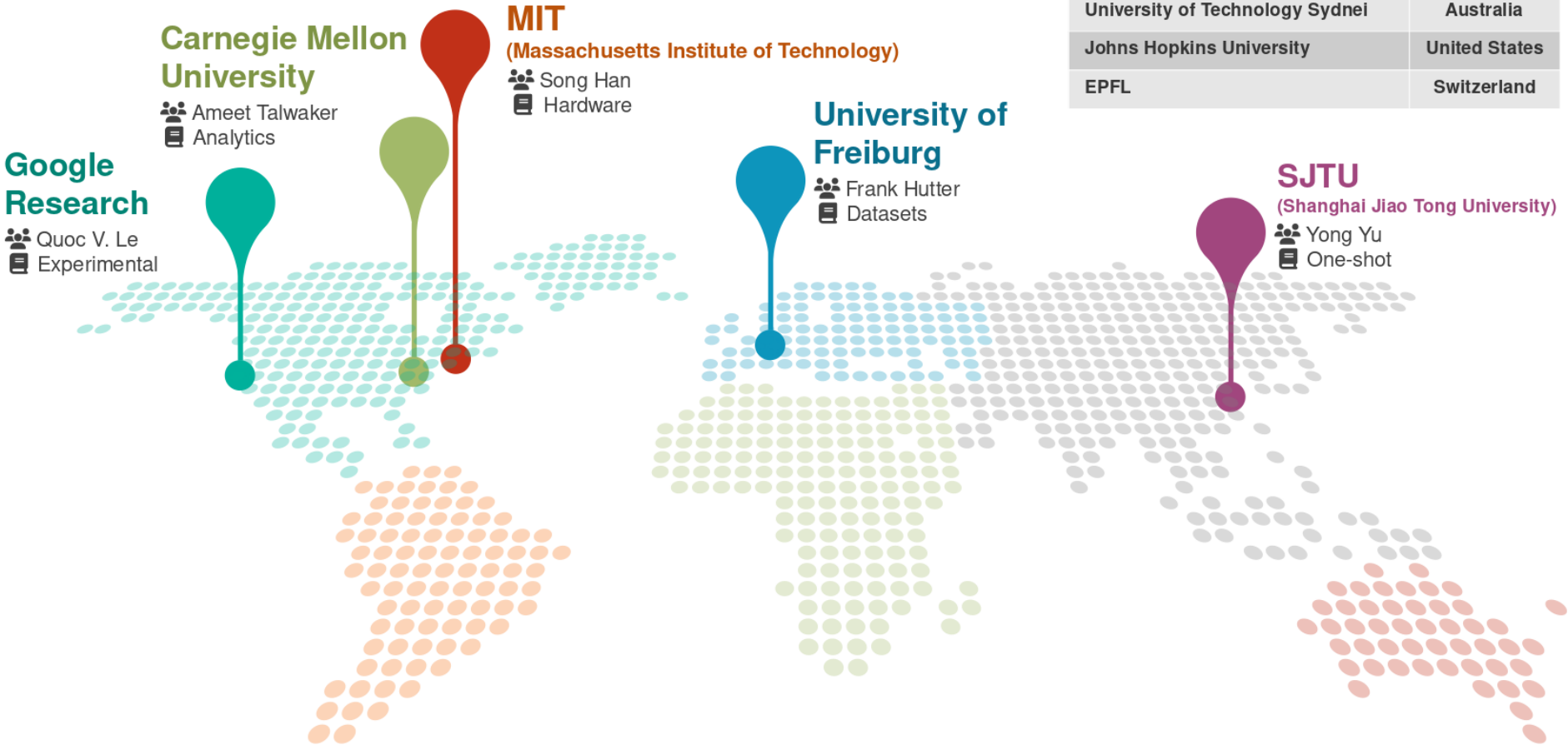
Agenda

Research areas



- **Hyperparameter Optimization** (HPO)
  - Tuning hyperparameters
  - Bayesian optimization
  - e.g.: Grid search
- **Meta-learning**
  - Reuse previous knowledge
  - e.g.: Transfer learning
  - e.g.: Few-shot learning
- **Neural Architecure Search** (NAS)
  - Search the network automatically

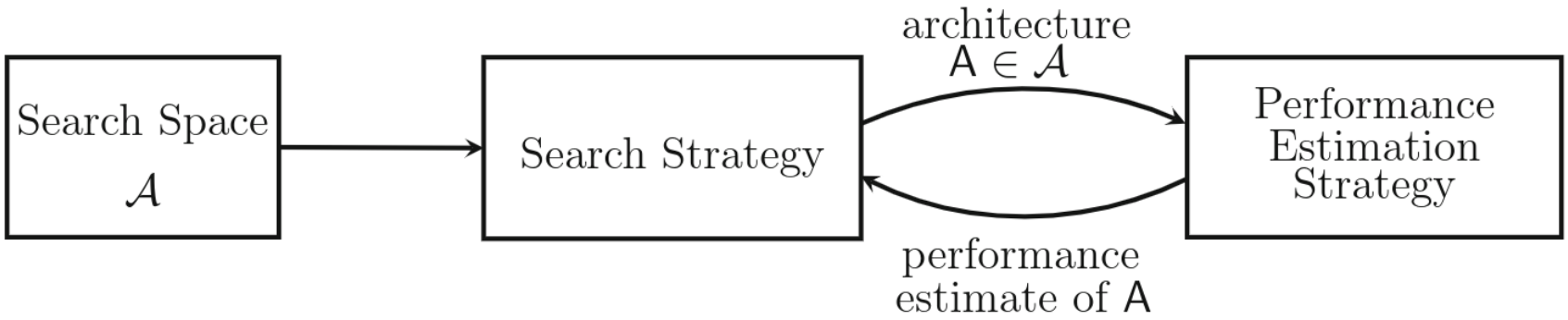
Research groups



Other groups:

Huawei Noah's Ark Lab	Various
Tongji University	China
University of Technology Sydnei	Australia
Johns Hopkins University	United States
EPFL	Switzerland

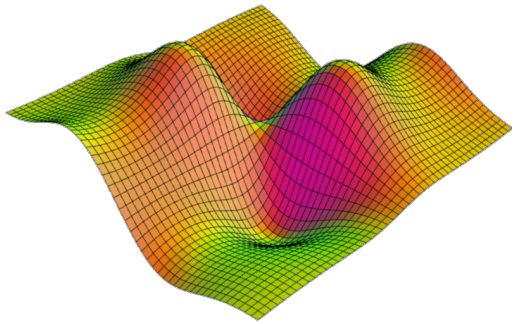
NAS framework

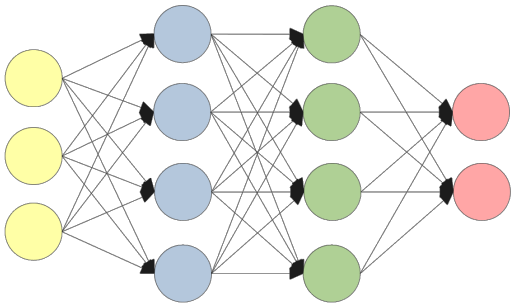


(Elsken, Metzen, and Hutter 2019)

NAS framework

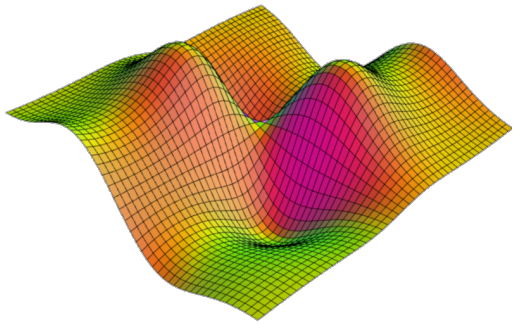
Search Space

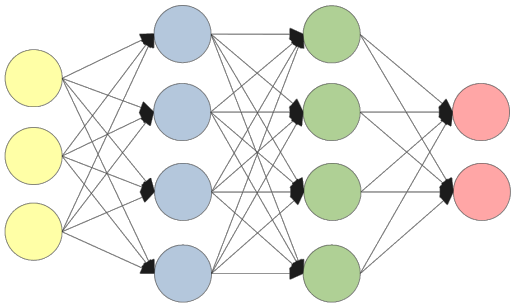




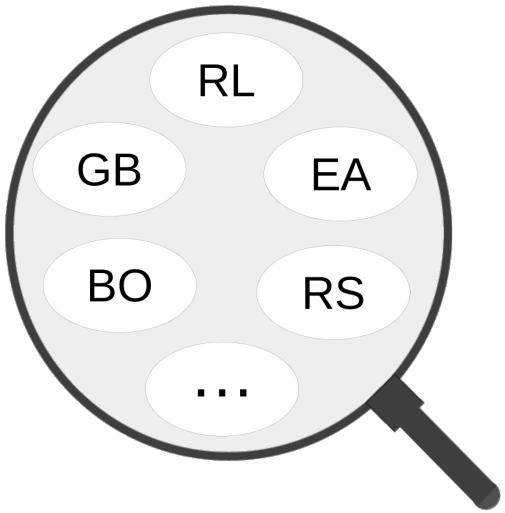
NAS framework

Search Space



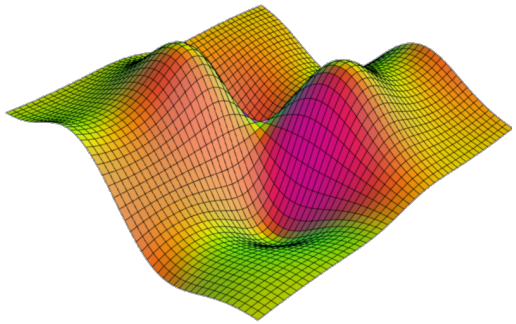


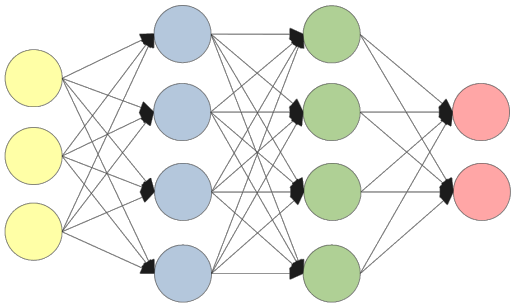
Search Strategy



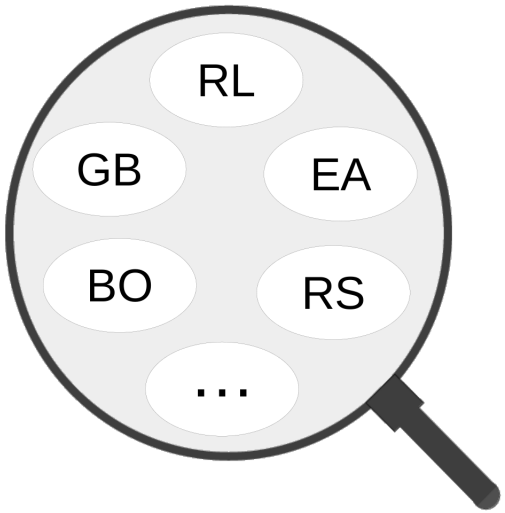
NAS framework

Search Space





Search Strategy



Search Evaluation

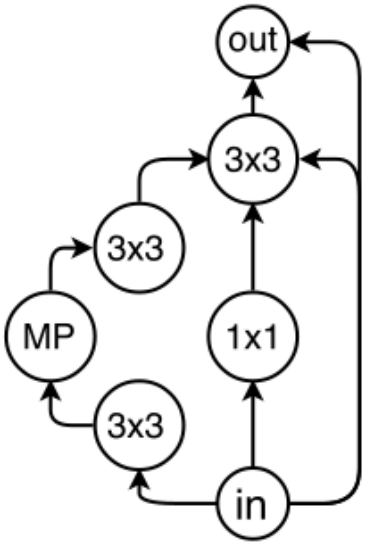
- Full training
- Partial training
- Weight-sharing
- Network morphism
- Hypernetworks
- Performance prediction

Search Space

Discrete Search Space:

- Vertices/Nodes: operations
- Edges: connections

Example:



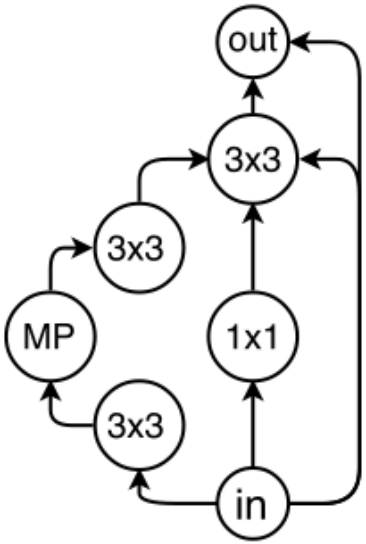
(Ying et al. 2019)

Search Space

Discrete Search Space:

- Vertices/Nodes: operations
- Edges: connections

Example:

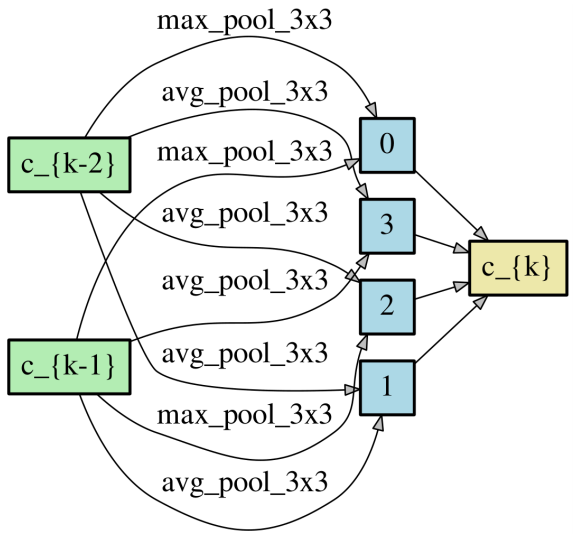


(Ying et al. 2019)

Continuous Search Space:

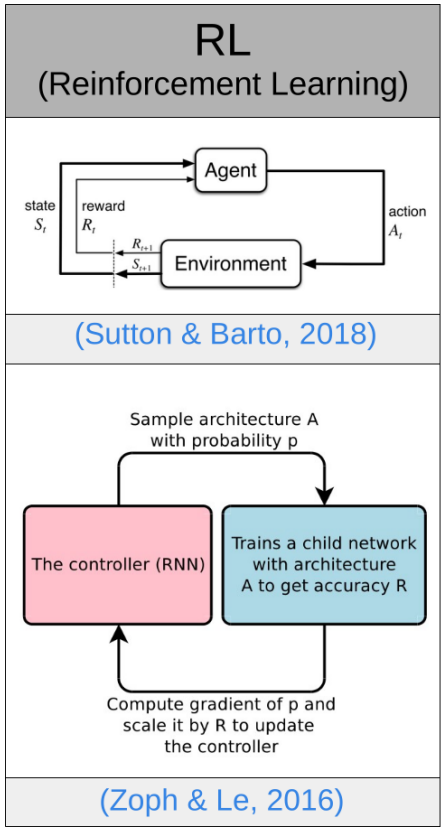
- Vertices/Nodes: latent representation (eg: feature map in conv nets)
- Edges: operations

Example:

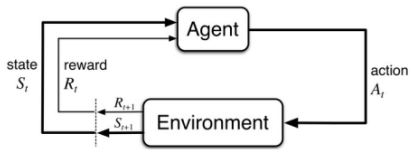
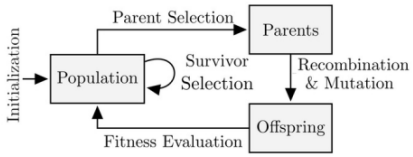
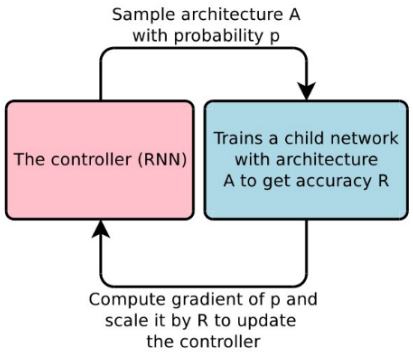
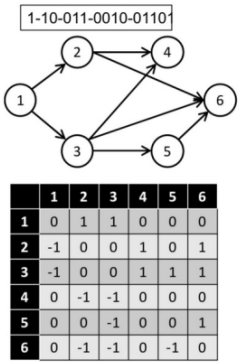


(Green et al. 2019)

Search Strategy

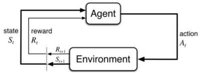
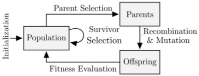
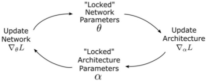
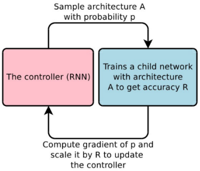
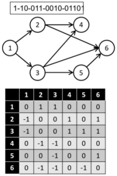
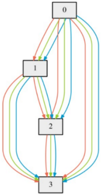


Search Strategy

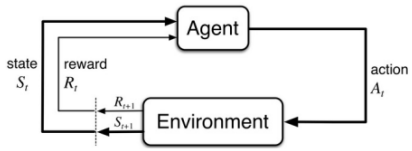
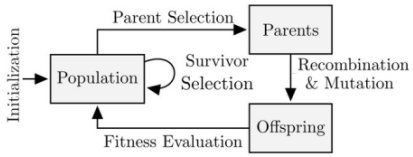
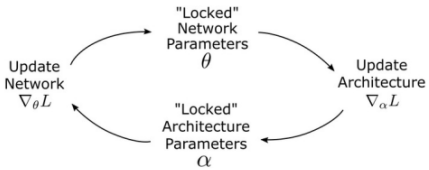
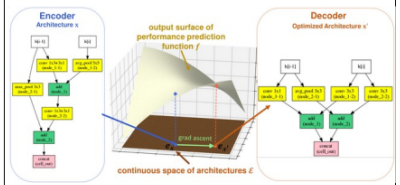
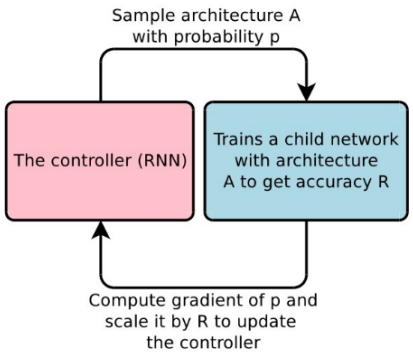
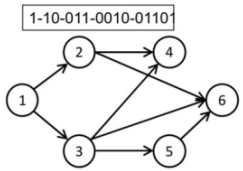
RL (Reinforcement Learning)	EA (Evolutionary Algorithms)
	
(Sutton & Barto, 2018)	(Wistuba et al., 2019)
	
(Zoph & Le, 2016)	(Lu et al., 2019)

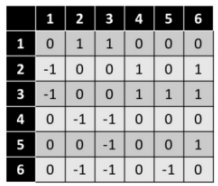
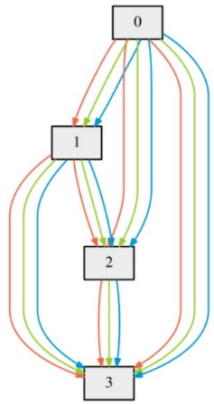
	1	2	3	4	5	6
1	0	1	1	0	0	0
2	-1	0	0	1	0	1
3	-1	0	0	1	1	1
4	0	-1	-1	0	0	0
5	0	0	-1	0	0	1
6	0	-1	-1	0	-1	0

Search Strategy

RL (Reinforcement Learning)	EA (Evolutionary Algorithms)	GB (Gradient-based)
		
(Sutton & Barto, 2018)	(Wistuba et al., 2019)	(Green et al., 2019)
		
(Zoph & Le, 2016)	(Lu et al., 2019)	(Liu et al., 2019b)

Search Strategy

<table>
<tr><th>RL
(Reinforcement Learning)</th><th>EA
(Evolutionary Algorithms)</th><th>GB
(Gradient-based)</th><th>Others</th></tr>
<tr><td> 
</td><td> 
</td><td> 
</td><td rowspan="2">• Bayesian Optimization
• Random Search
• Model-based
◦ Gaussian process
◦ Random Forest
◦ XGBoost
  </td></tr>
<tr><td>(Sutton & Barto, 2018)</td><td>(Wistuba et al., 2019)</td><td>(Green et al., 2019)</td></tr>
<tr><td> 
</td><td> 


</td><td> 
</td></tr>
<tr><td>(Zoph & Le, 2016)</td><td>(Lu et al., 2019)</td><td>(Liu et al., 2019b)</td><td>(Luo et al., 2019)</td></tr>
</table>

NAS framework

- **Search Space**:
  - Cell-based
  - One-Shot
- **Search Strategy**:
  - Reinforcement Learning
  - Gradient-based
  - Evolutionary Algorithms
- **Performance Estimation**:
  - Full training
  - Partial training
  - Performance prediction

Once-for-All

- **Search Space**:
  - Cell-based
  - One-Shot $\longrightarrow$ Training stage
- **Search Strategy**:
  - Reinforcement Learning
  - Gradient-based $\longrightarrow$ Training stage
  - Evolutionary Algorithms $\longrightarrow$ Search stage
- **Performance Estimation**:
  - Full training
  - Partial training
  - Performance prediction $\longrightarrow$ Search stage

Agenda

OFA network

- Convolutional Neural Network (CNN)
- Same architecture space of MobileNetV3
- Supernetwork with smaller subnetworks nested inside larger architectures

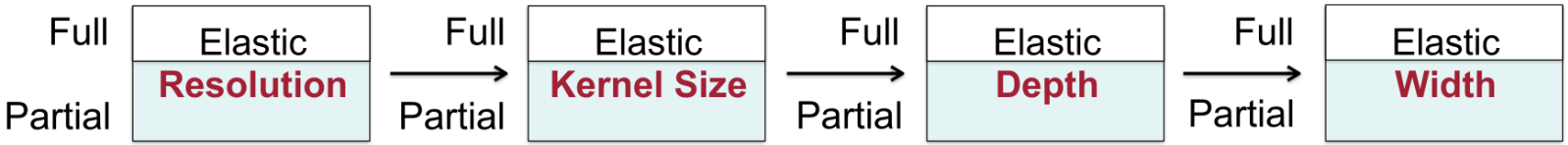
- Final architecture: 5 units
- Each unit have 3 adjustable parameters
  - depth $\in \{2, 3, 4\}$
  - width $\in \{3, 4, 6\}$
  - kernel size $\in \{3, 5, 7\}$
- The 4_{th} parameter is unique for the architecture
  - image resolution $\in \{128, 132, \ldots, 224\}$

OFA training

- Start training the network at its full size
  - Optimizer = SGD
  - Nesterov momentum = 0.9
  - Weight decay = $3e^{-5}$
  - Initial learning rate = 2.6
  - Learning weight decay = cosine schedule
  - Epochs = 180
  - Batch size = 2048
  - GPUs = 32 V100 GPUs
  - Cost = 1,200 GPU hours

(Cai et al. 2020).

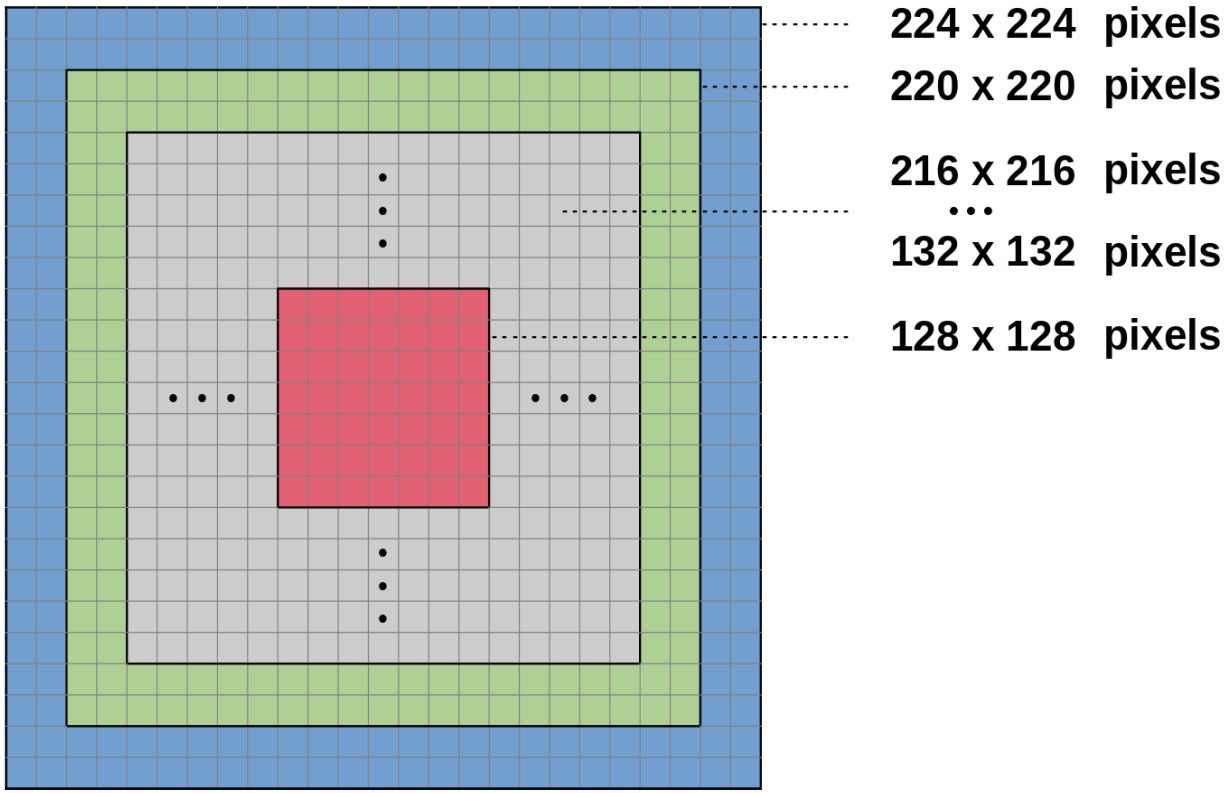
Progressive shrinking



- Progressively shrinks the network using the *Progressive shrinking* algorithm

Elastic resolution

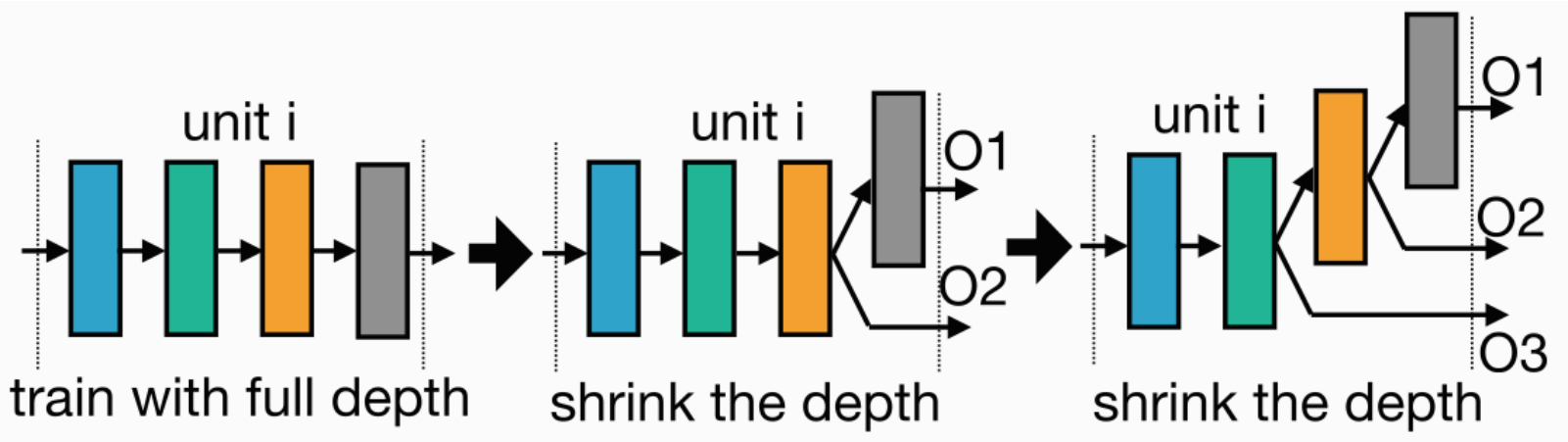
- 25 different input resolutions: {128, 132, . . . , 224}



(Cai et al. 2020)

Elastic depth

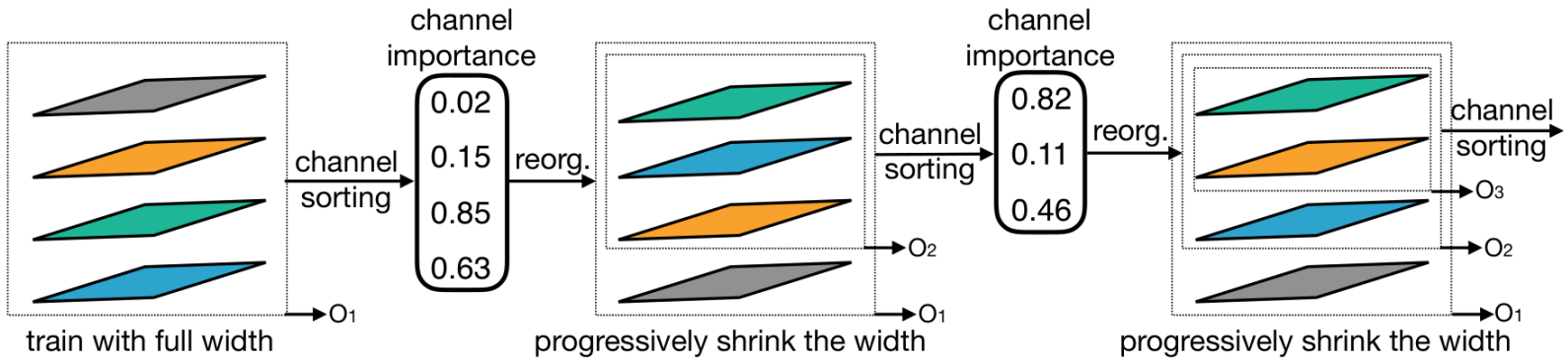
- Depth options: {2, 3, 4} layers.
- Start training with 4 layers, then trains with 3 layer, and lastly with 2 layers.



(Cai et al. 2020)

Elastic width

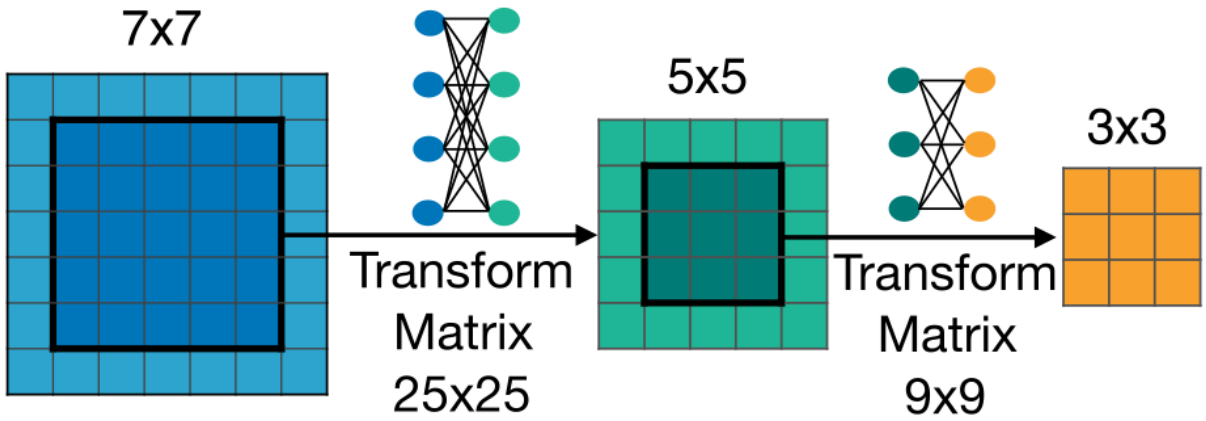
- Channel expantion ratio: {3, 4, 6}.
- Start training with all filters. Then skip the least important one by one.
- The importance of each channel is calculated using the L1 norm of the channel's weights.



(Cai et al. 2020)

Elastic kernel size

- Convolutional kernel size: {7x7, 5x5, 3x3}
- Start training with kernel 7x7, then 5x5, and lastly 3x3.



(Cai et al. 2020)

- Different layers have different kernel transformation matrices.
- The matrices are shared among different channels within each layer.
- Total kernel transformation parameters: $25 \times 25 + 9 \times 9 = 706$ per layer

Agenda

Performance estimator

- **Accuracy Predictor**
  - Feedforward neural network with
  - 3 layers
  - 400 hidden units per layer
- **Efficiency Predictor**
  - Latency lookup table (Cai, Zhu, and Han 2022)
  - Function to count FLOPS given the operations.

Architecture Space

PS	property	options	# choices
elastic resolution	input resolution	$\{224, \ldots, 128\}$	25
elastic kernel	kernel size	$\{3, 5, 7\}$	3
elastic depth	layers	$\{2, 3, 4\}$	3
elastic width	channels	$\{3, 4, 6\}$	3

Architecture Space

PS	property	options	# choices
elastic resolution	input resolution	{224, . . . , 128}	25
elastic kernel	kernel size	{3, 5, 7}	3
elastic depth	layers	{2, 3, 4}	3
elastic width	channels	{3, 4, 6}	3

```
{'ks': [5, 7, 3, 5, 5, 7, 7, 5, 3, 3, 3, 3, 7, 5, 3, 7, 7, 5, 7, 5],
 'e': [4, 4, 4, 3, 3, 4, 3, 4, 6, 6, 4, 3, 6, 3, 6, 4, 3, 3, 4, 3],
 'd': [2, 4, 2, 3, 2],
 'r': [176]}
```

Architecture Space

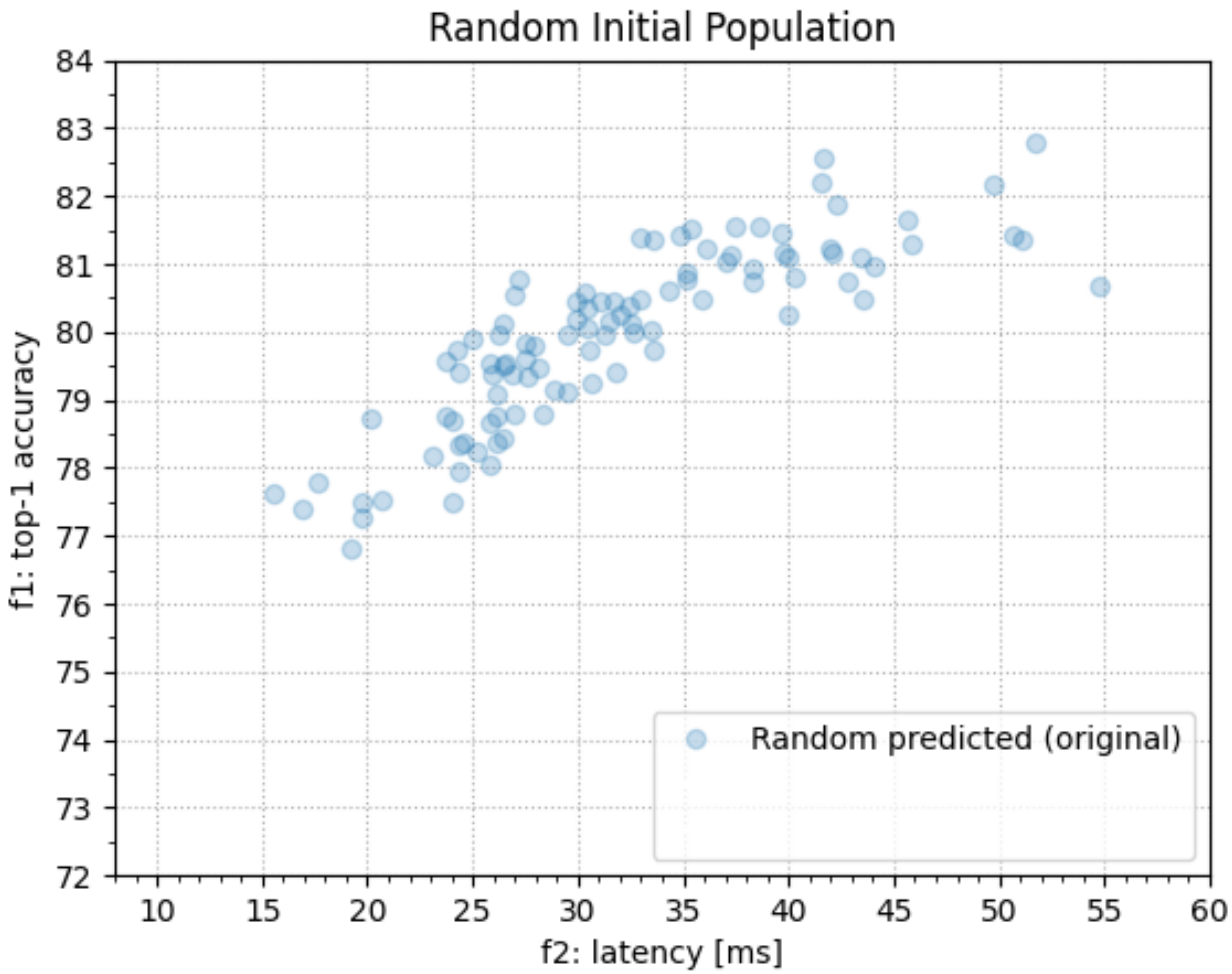
PS	property	options	# choices
elastic resolution	input resolution	{224, . . . , 128}	25
elastic kernel	kernel size	{3, 5, 7}	3
elastic depth	layers	{2, 3, 4}	3
elastic width	channels	{3, 4, 6}	3

```
{'ks': [5, 7, 3, 5, 5, 7, 7, 5, 3, 3, 3, 3, 7, 5, 3, 7, 7, 5, 7, 5],
 'e': [4, 4, 4, 3, 3, 4, 3, 4, 6, 6, 4, 3, 6, 3, 6, 4, 3, 3, 4, 3],
 'd': [2, 4, 2, 3, 2],
 'r': [176]}
```

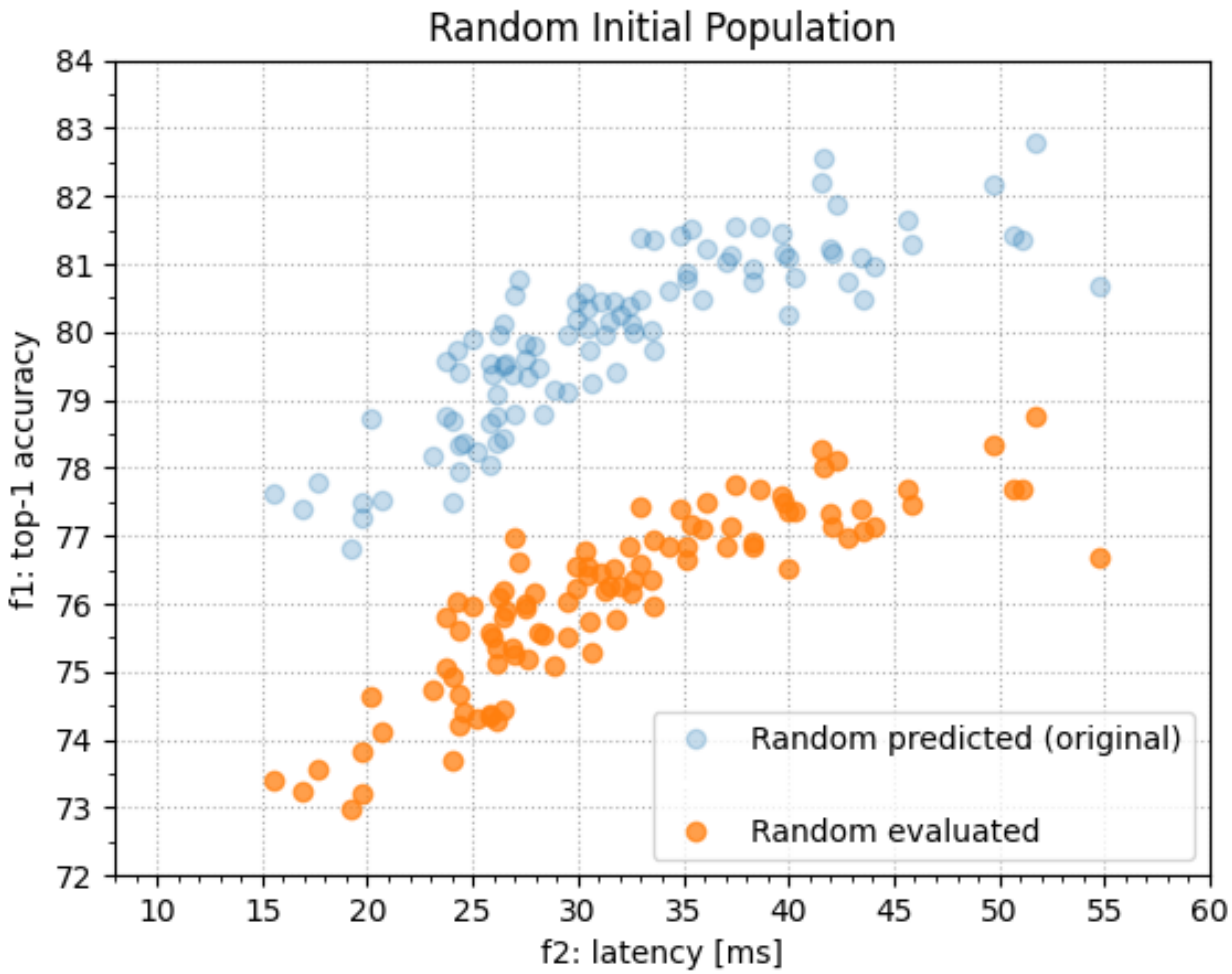
- Search space of $((3\times3)^2 + (3\times3)^3 + (3\times3)^4)^5 \approx 2\times10^{19}$ different neural network architectures.
- All these subnetworks share the same weights: total of 7.7M parameters.

Agenda

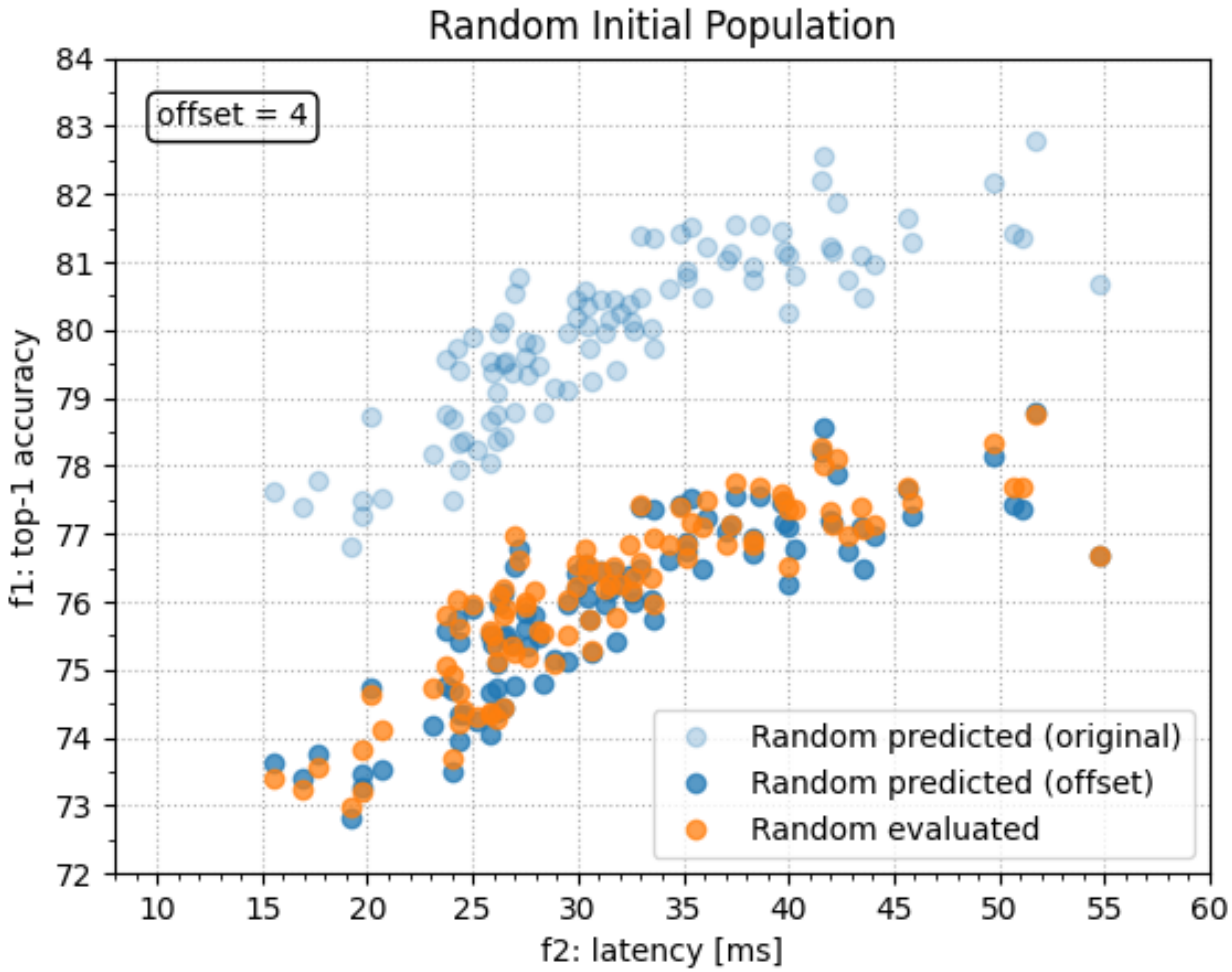
Random search



Random search

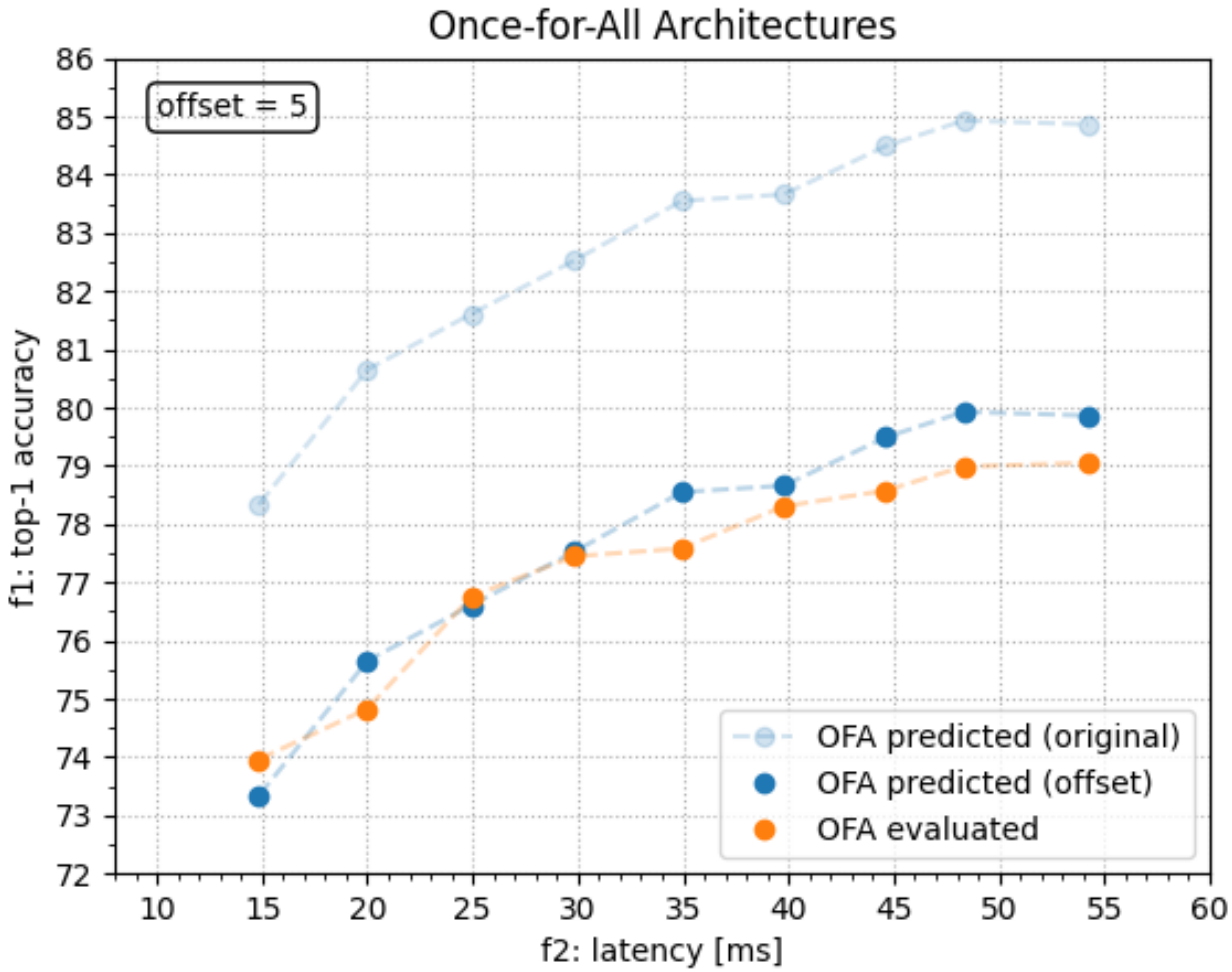


Random search



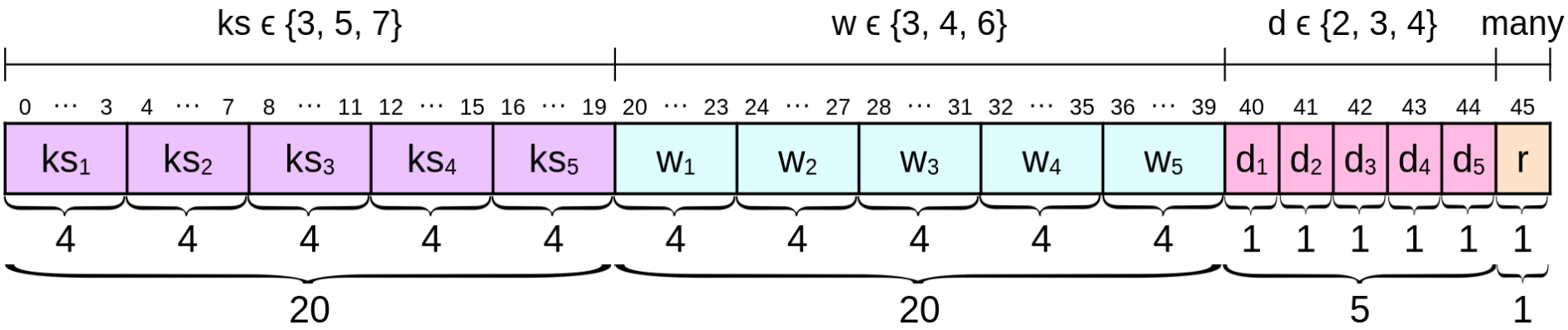
Agenda

OFA search

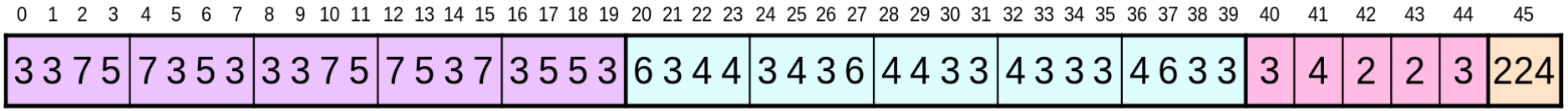

Once-for-All Architectures
offset = 5
f1: top-1 accuracy
f2: latency [ms]
OFA predicted (original)
OFA predicted (offset)
OFA evaluated
86
85
84
83
82
81
80
79
78
77
76
75
74
73
72
10
15
20
25
30
35
40
45
50
55
60

Agenda

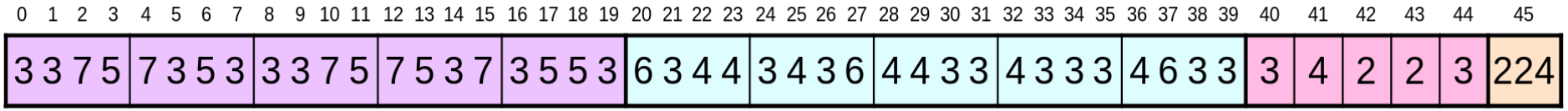
Encoding


ks ϵ {3, 5, 7}
w ϵ {3, 4, 6}
d ϵ {2, 3, 4}
many
0 ⋯ 3
4 ⋯ 7
8 ⋯ 11
12 ⋯ 15
16 ⋯ 19
20 ⋯ 23
24 ⋯ 27
28 ⋯ 31
32 ⋯ 35
36 ⋯ 39
40
41
42
43
44
45
ks1
ks2
ks3
ks4
ks5
w1
w2
w3
w4
w5
d1
d2
d3
d4
d5
r
4
4
4
4
4
4
4
4
4
4
1
1
1
1
1
1
20
20
5
1

Encoding example



Encoding example



																	size	index
ks	3	3	7	5	7	3	5	3	3	3	7	5	7	5	3	7	3	5	5	3	20	[00:19]
w	6	3	4	4	3	4	3	6	4	4	3	3	4	3	3	3	4	6	3	3	20	[20:39]
d	3				4				2				2				3				5	[40:44]
r	224																				1	[45]

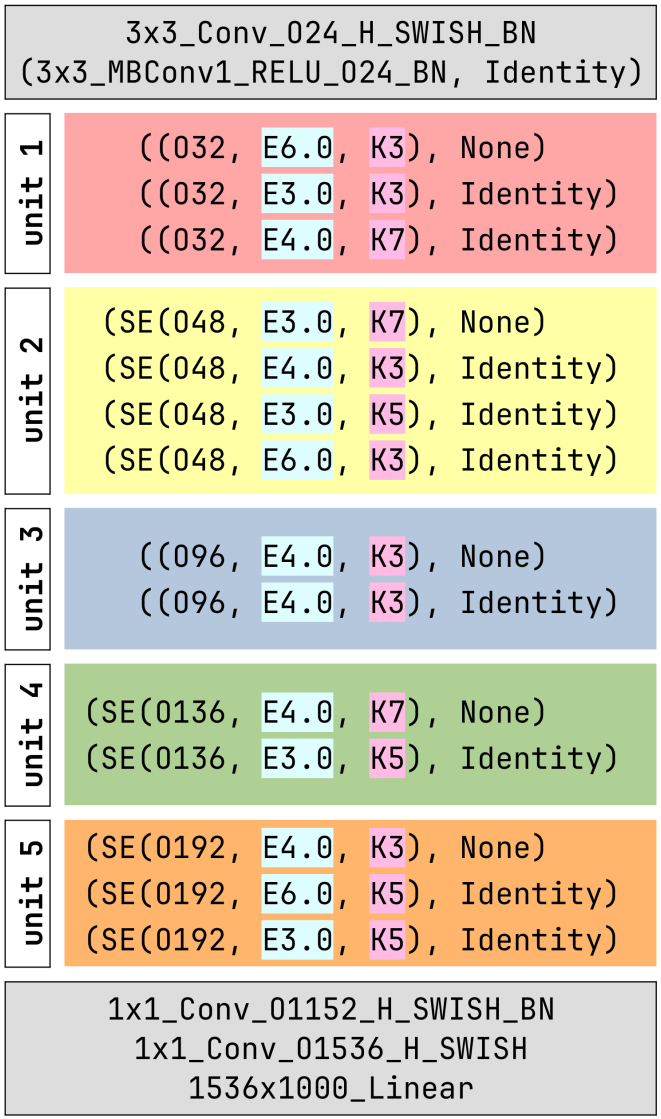
Encoding example

0 1 2 3	4 5 6 7	8 9 10 11	12 13 14 15	16 17 18 19	20 21 22 23	24 25 26 27	28 29 30 31	32 33 34 35	36 37 38 39	40	41	42	43	44	45
3 3 7 5	7 3 5 3	3 3 7 5	7 5 3 7	3 5 5 3	6 3 4 4	3 4 3 6	4 4 3 3	4 3 3 3	4 6 3 3	3	4	2	2	3	224

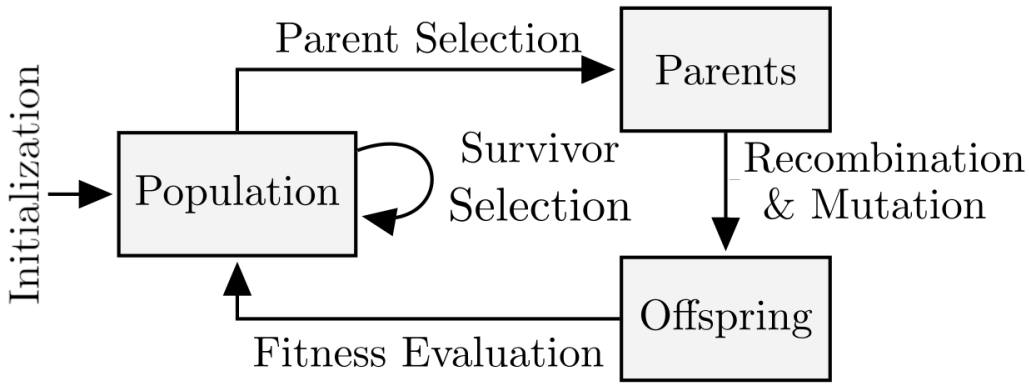
											size	index
ks	3	3	7	5	7	3	5	3	3	3	7	5	7	5	3	7	3	5	5	3	20	[00:19]
w	6	3	4	4	3	4	3	6	4	4	3	3	4	3	3	3	4	6	3	3	20	[20:39]
d	3				4				2				2				3				5	[40:44]
r	224																				1	[45]

	Unit 1				Unit 2				Unit 3				Unit 4				Unit 5				size	index
ks	3	3	7	5	7	3	5	3	3	3	7	5	7	5	3	7	3	5	5	3	20	[00:19]
w	6	3	4	4	3	4	3	6	4	4	3	3	4	3	3	3	4	6	3	3	20	[20:39]
d	3				4				2				2				3				5	[40:44]
r	224																				1	[45]

Encoding example



Evolutionary Algorithm framework

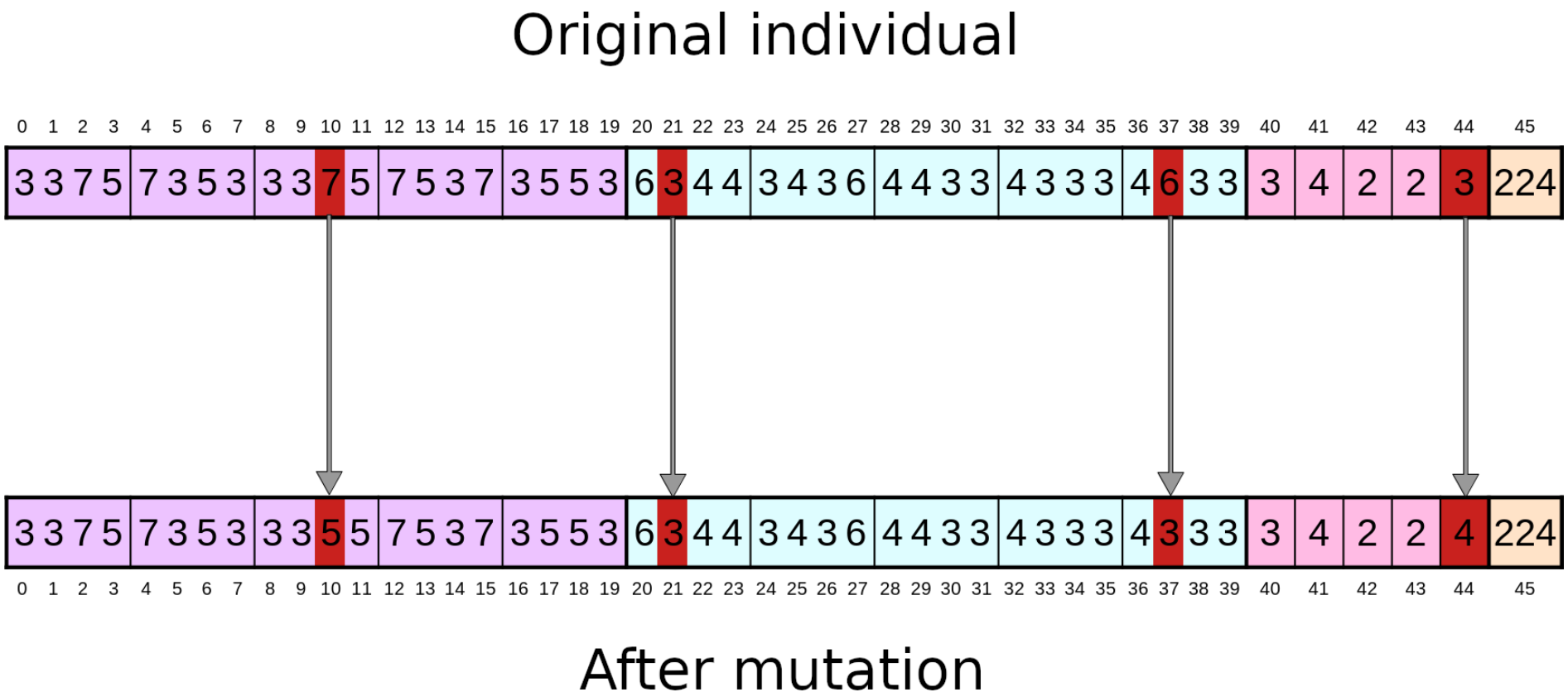


(Real et al. 2017)

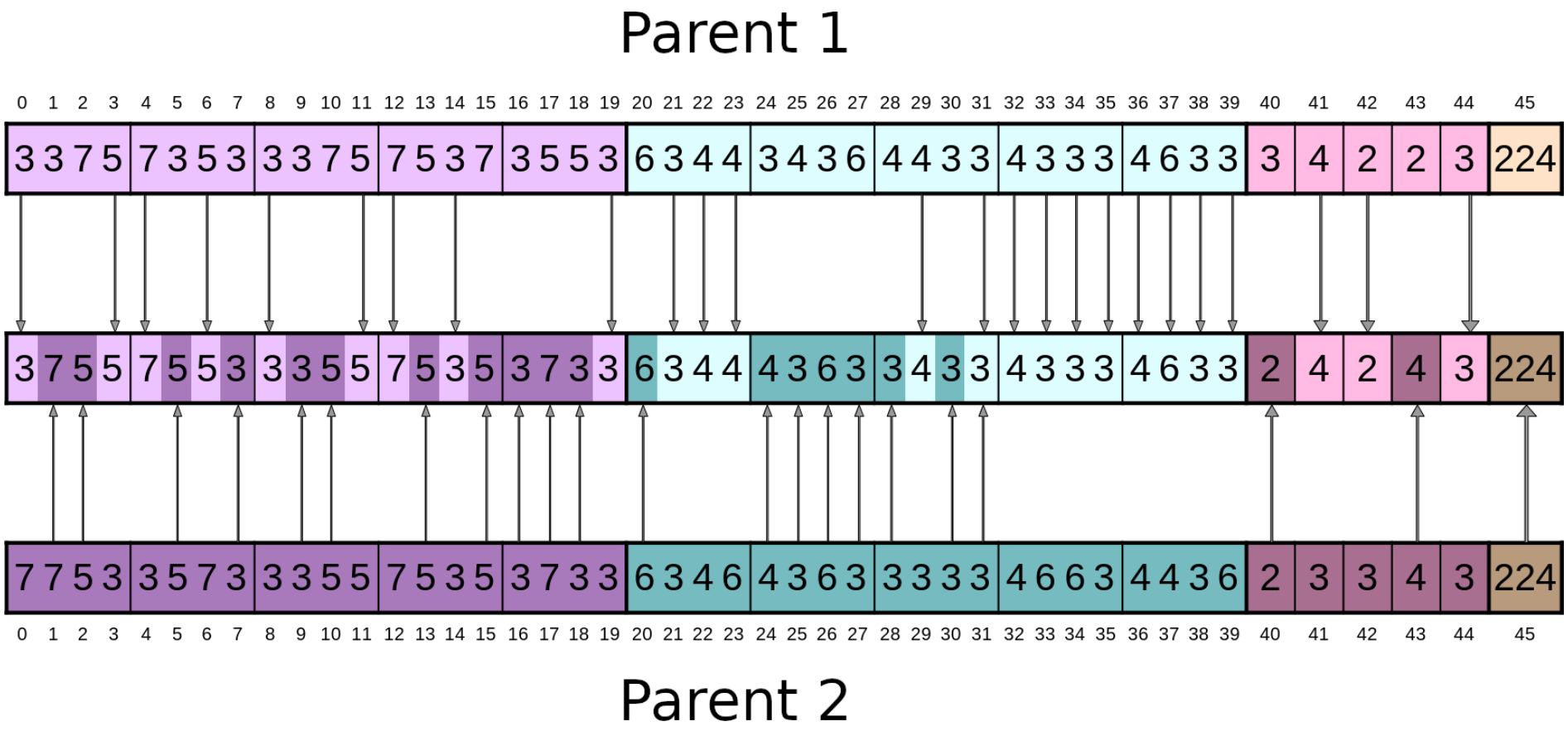
Elements:

- **Individual**: candidate solution
- **Encoding**: representation of a candidate solution
- **Fitness function**: evaluation of the candidate solution
- **Population**: group of individuals
- **Operators**:
  - **Sampling**
  - **Mutation**
  - **Crossover**
  - **Selection**
- **Offspring**: next generation population

Operator: Mutation



Operator: Crossover (recombination)

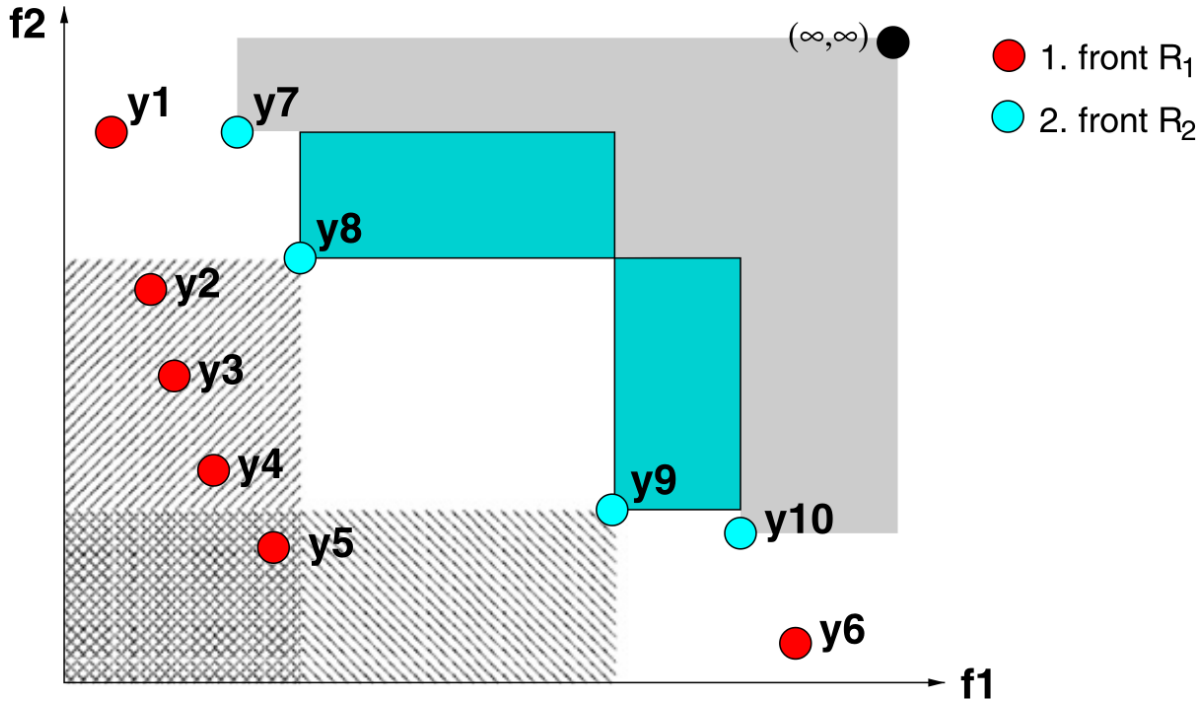


Hyperparameters

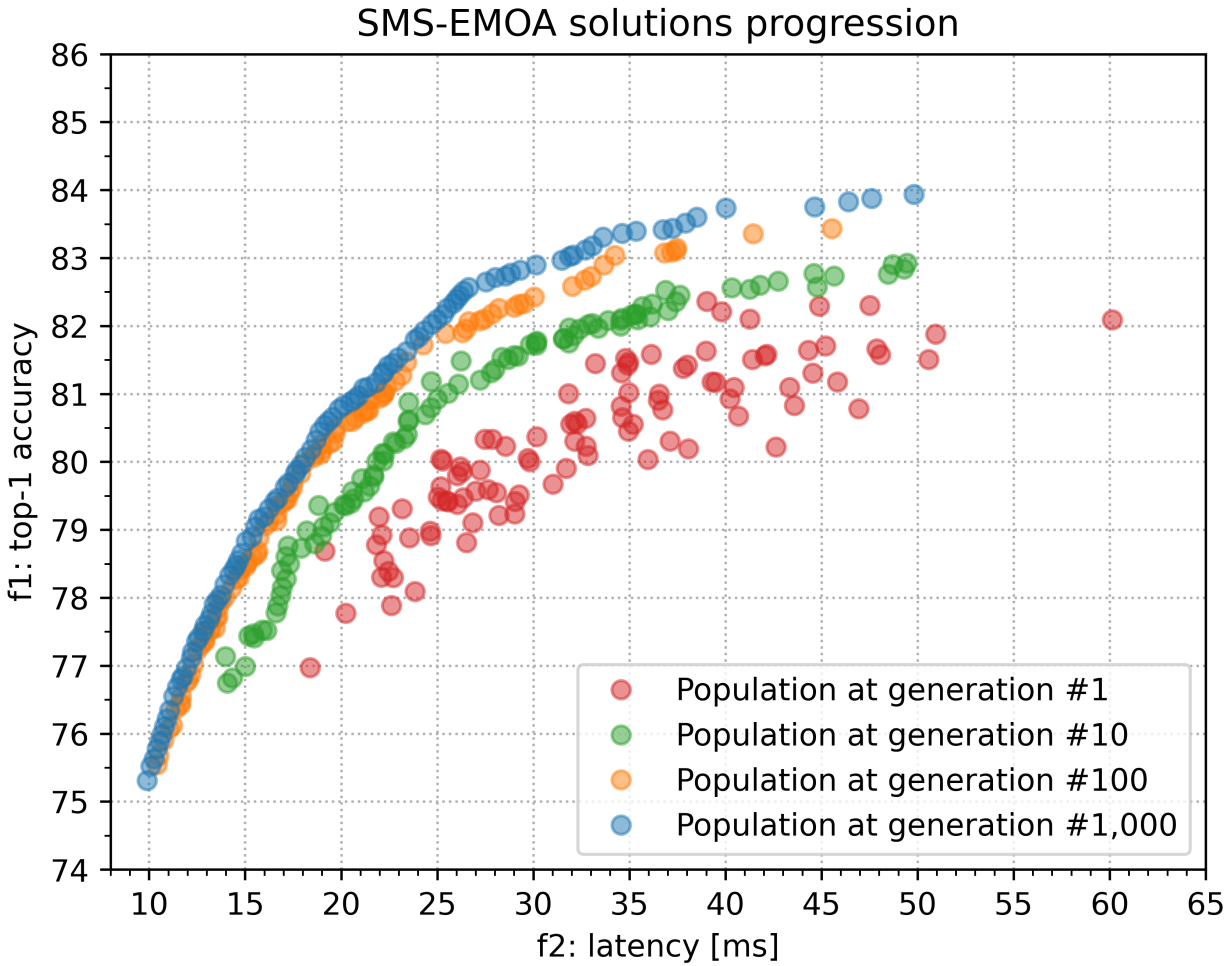
- Population: 100
- Generations: 1,000
- Operators:
  - Mutation: 50% of population
    - Mutation architecture (bitflip): 10%
  - Crossover: 25% as parents
    - Uniform
- Algorithms:
  - NSGA-II
  - SMS-EMOA
- Fitness function:
  - Accuracy predictor
  - Efficiency predictor
    - Latency
    - FLOPS

SMS-EMOA

- Maximizes the hypervolume

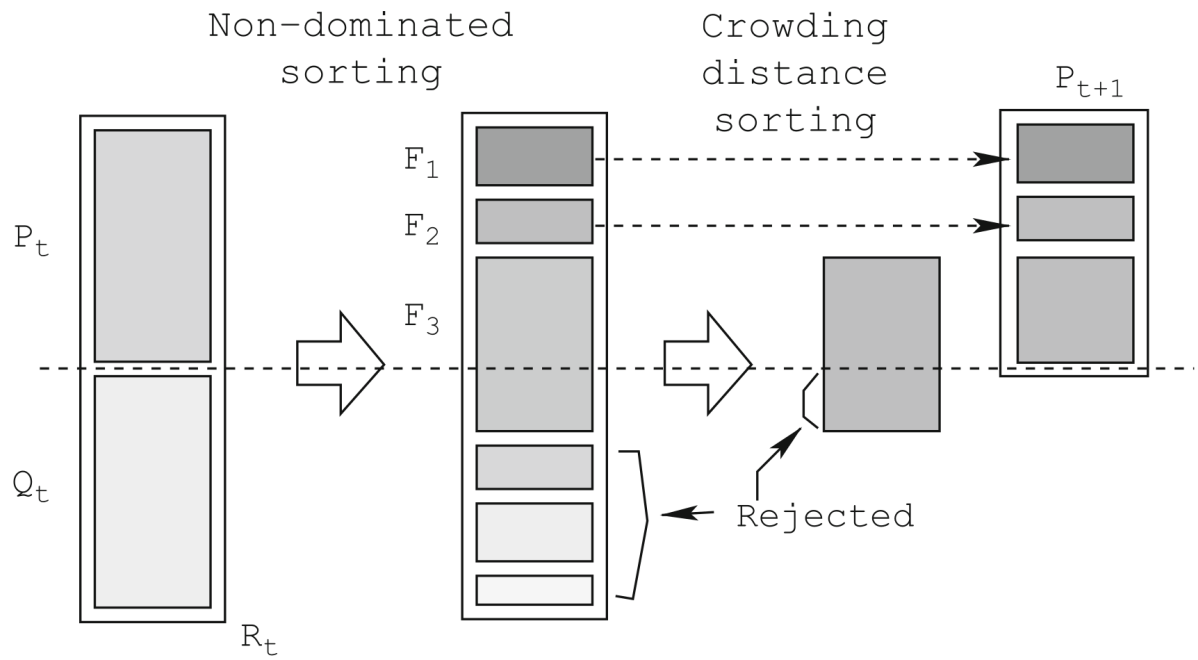


SMS-EMOA: Population

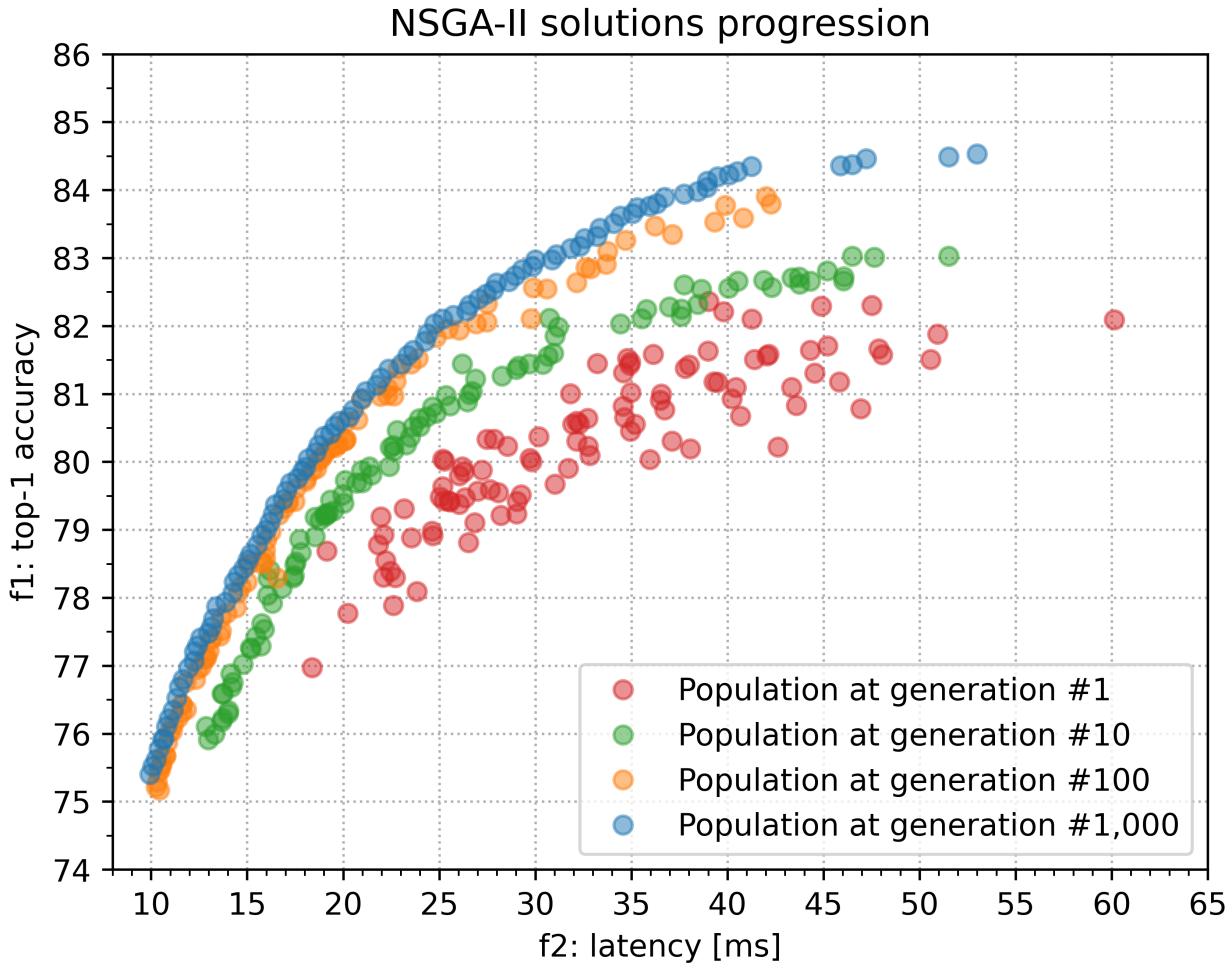


NSGA-II

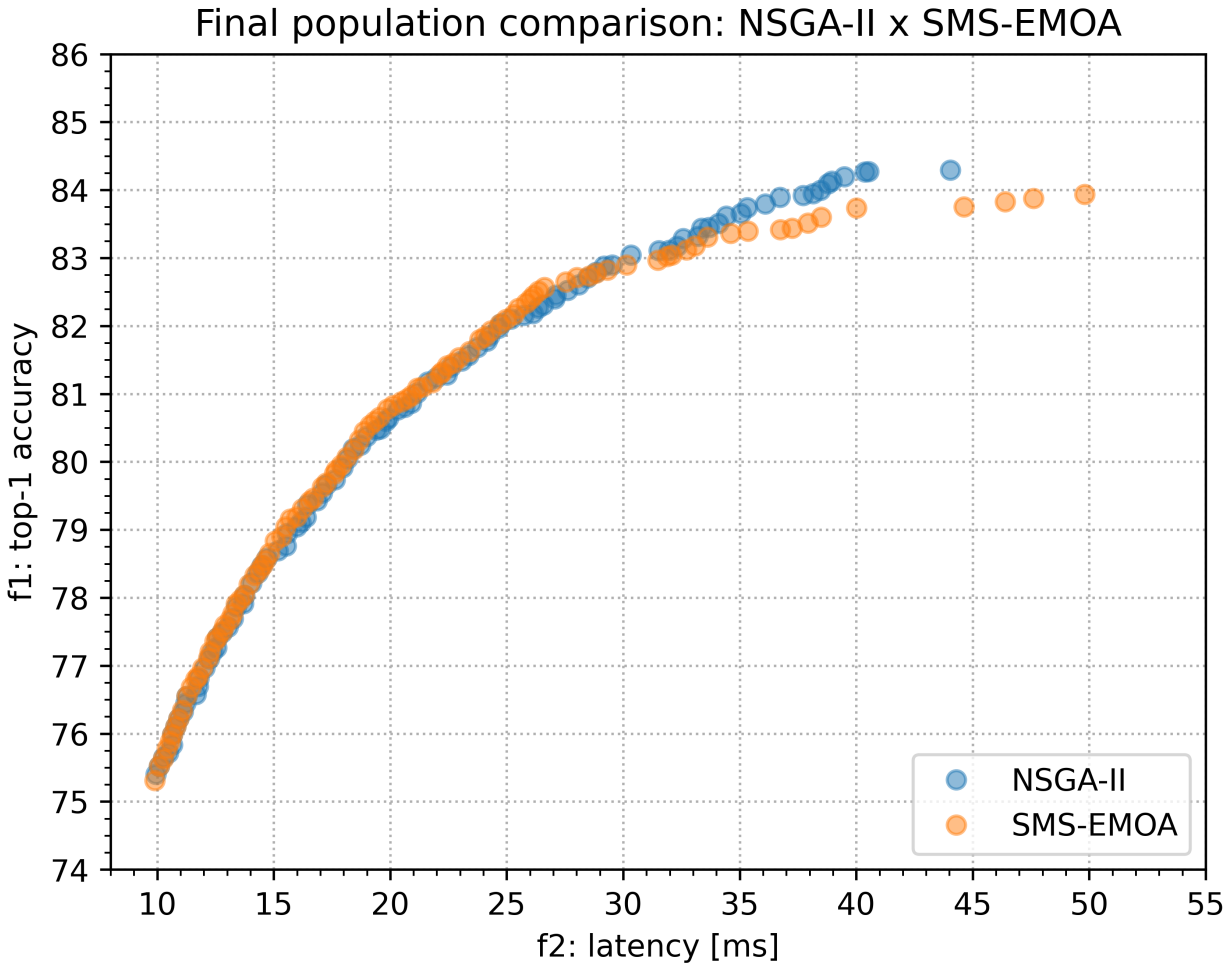
- **Sorting**:
  - Non-dominated
  - Crowding distance



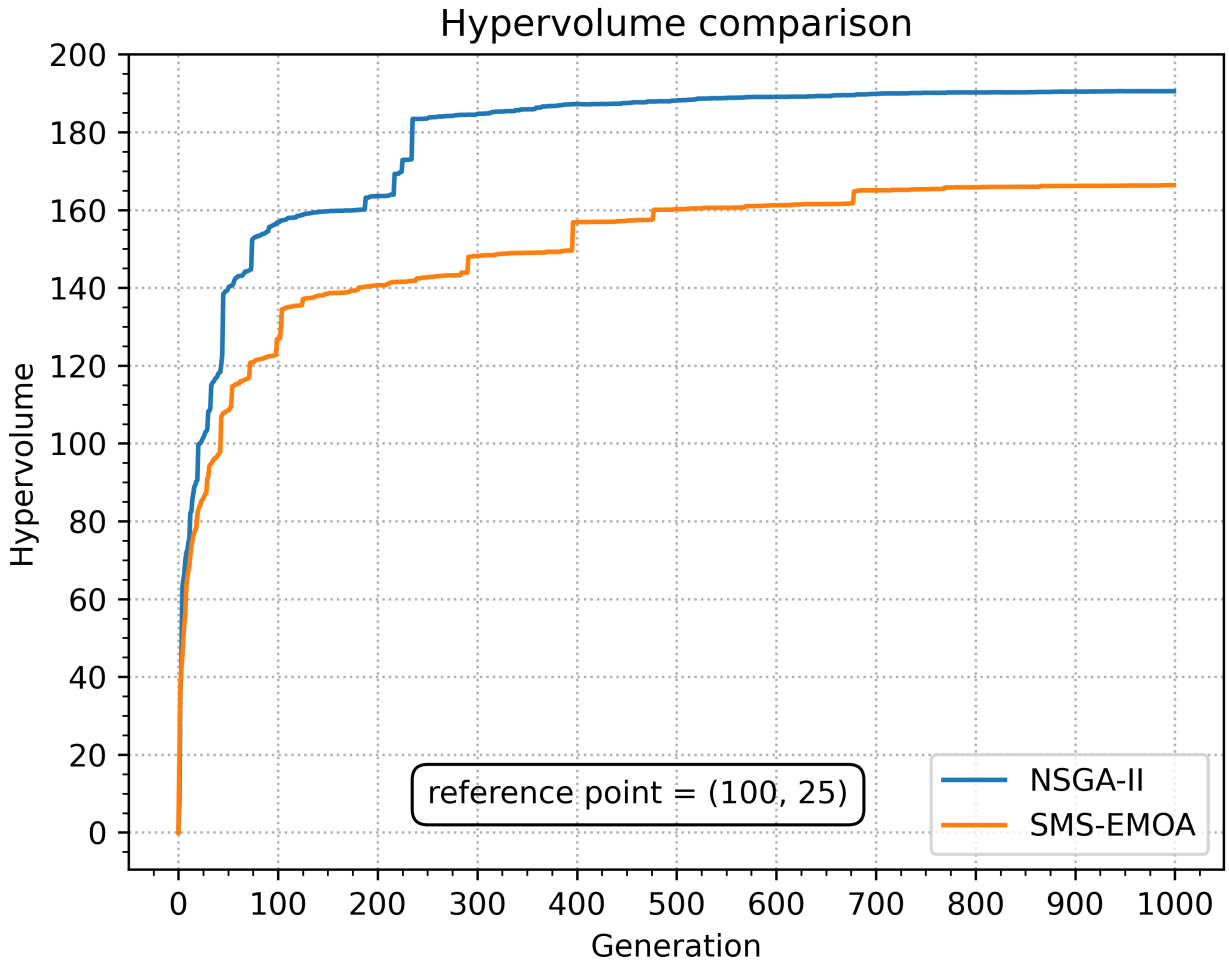
NSGA-II: Population



Population comparison

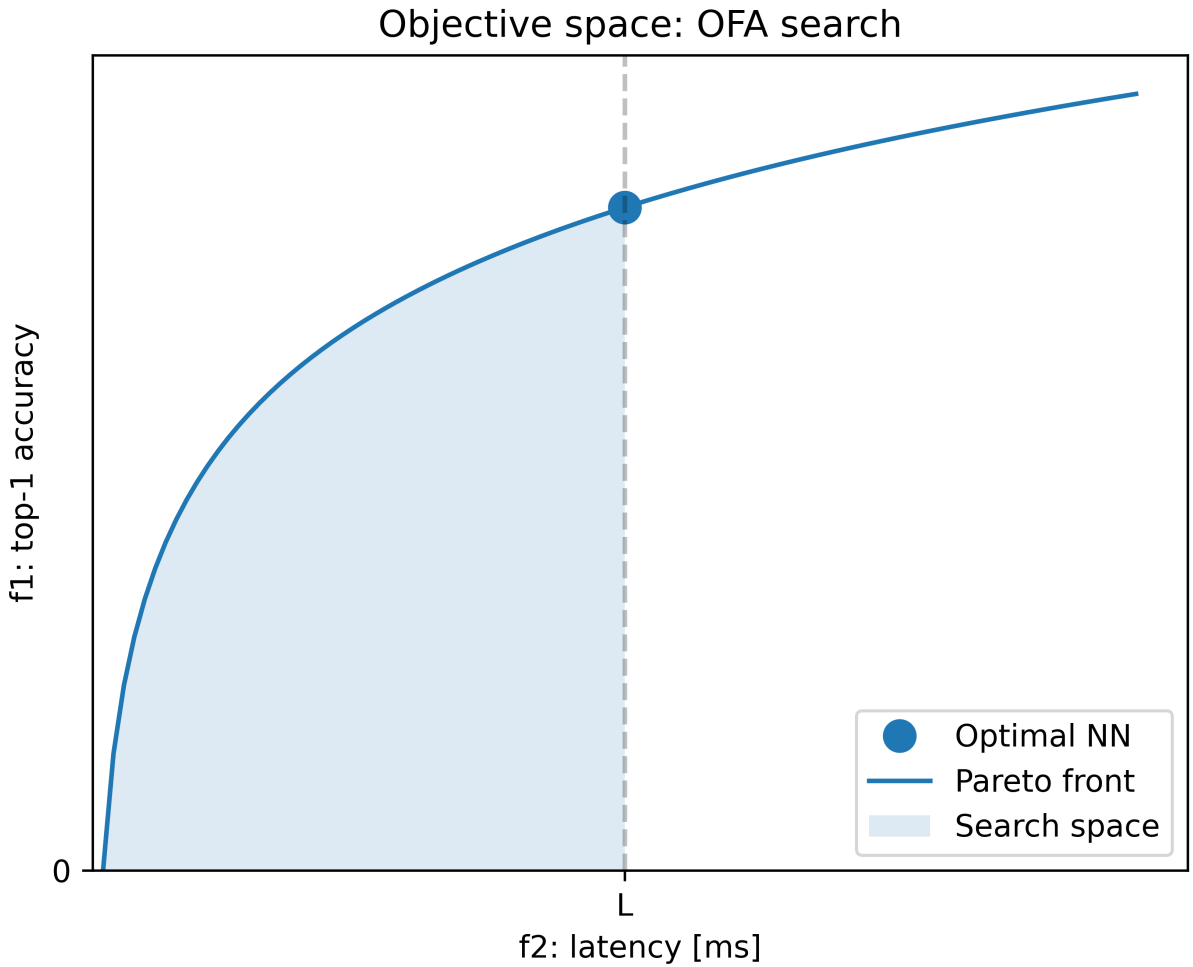


Hypervolume

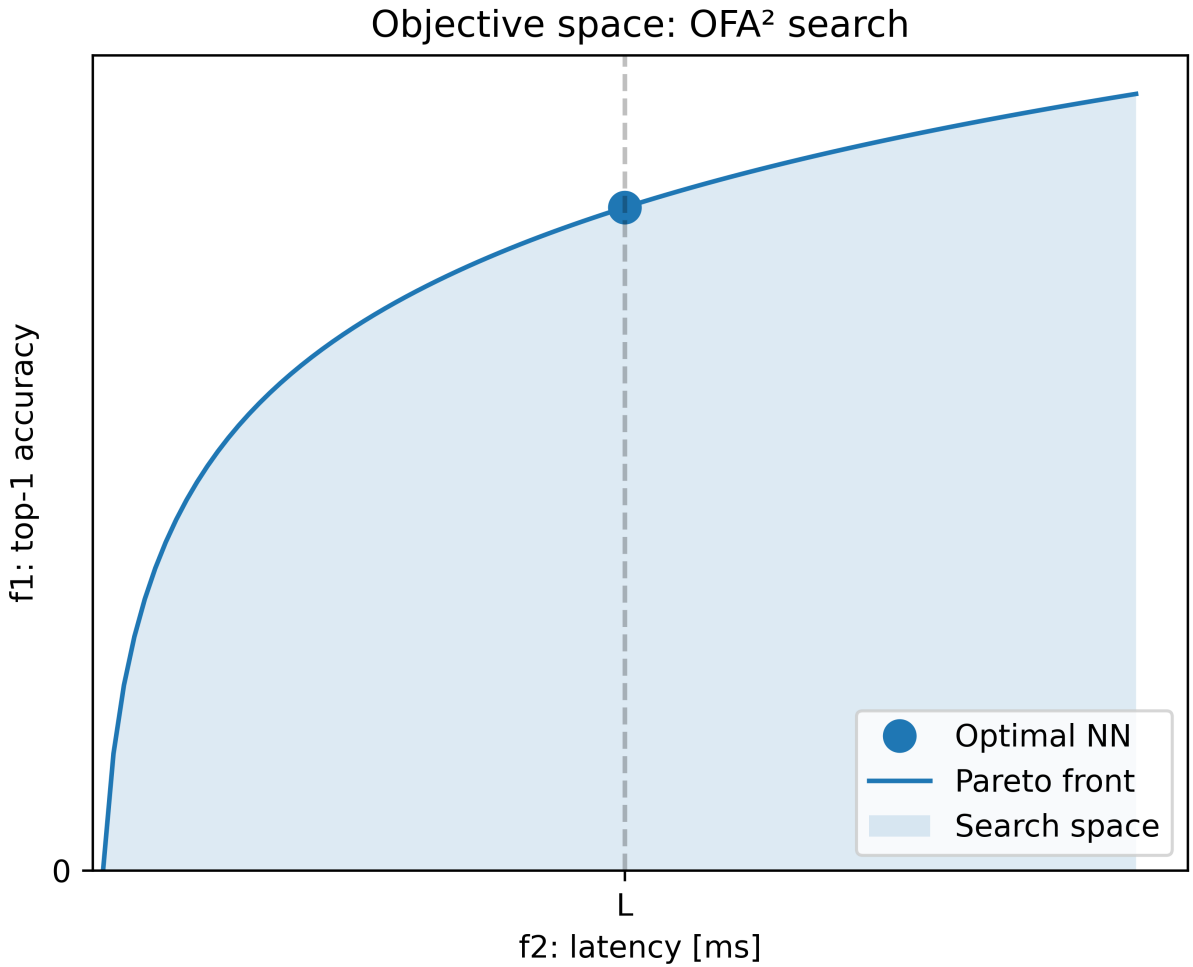


Agenda

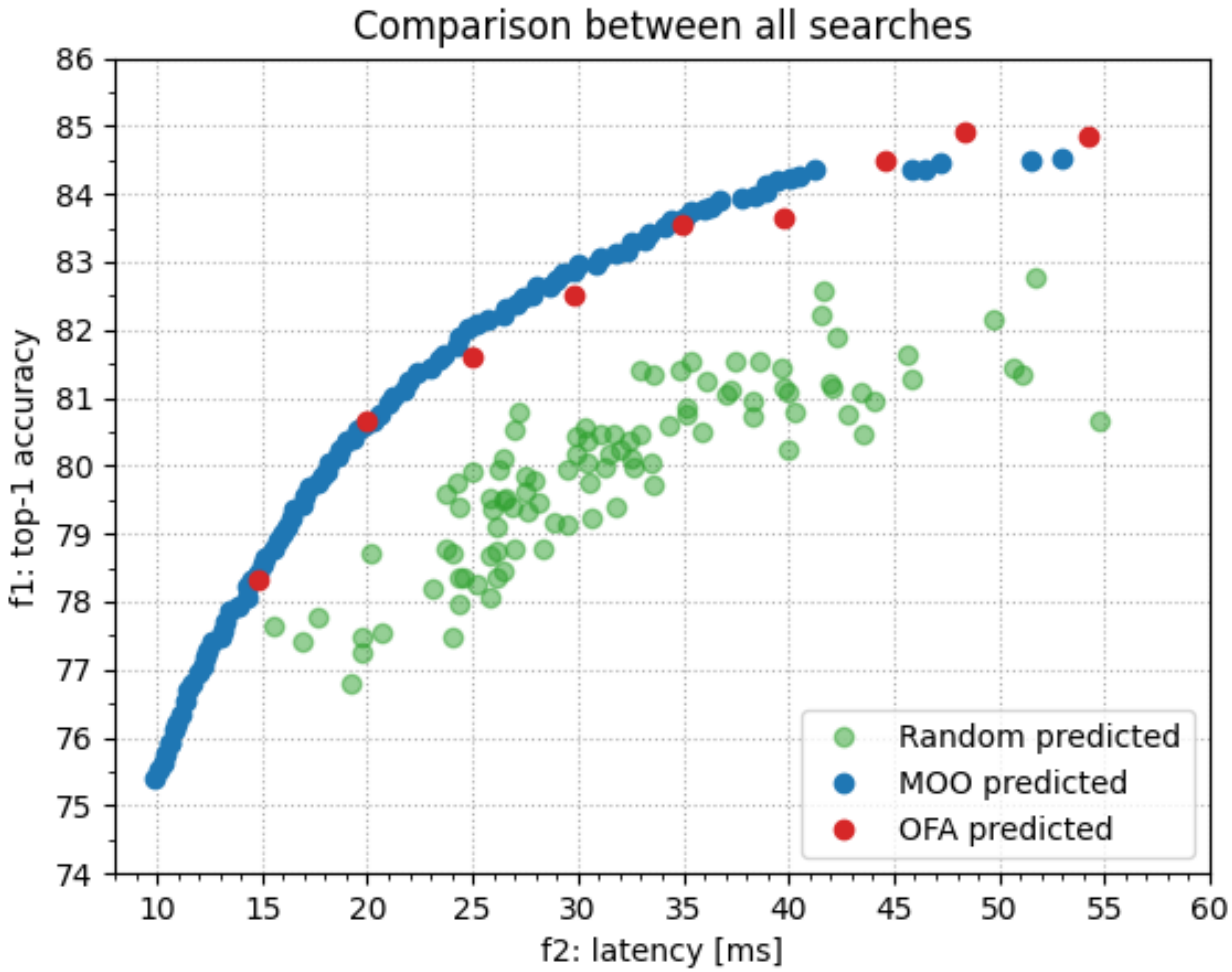
OFA search space



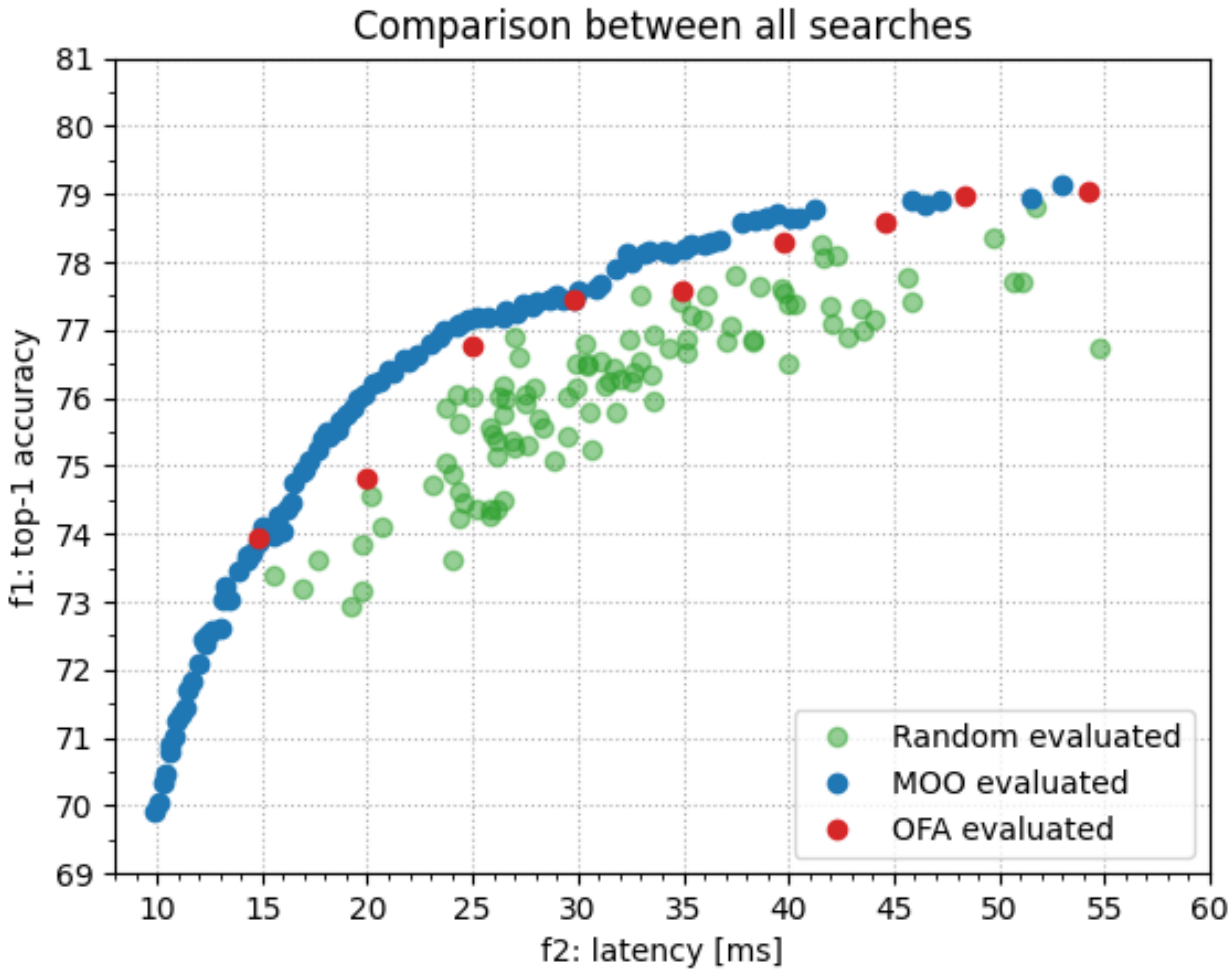
OFA² search space



Predicted accuracy



Evaluated accuracy



Agenda

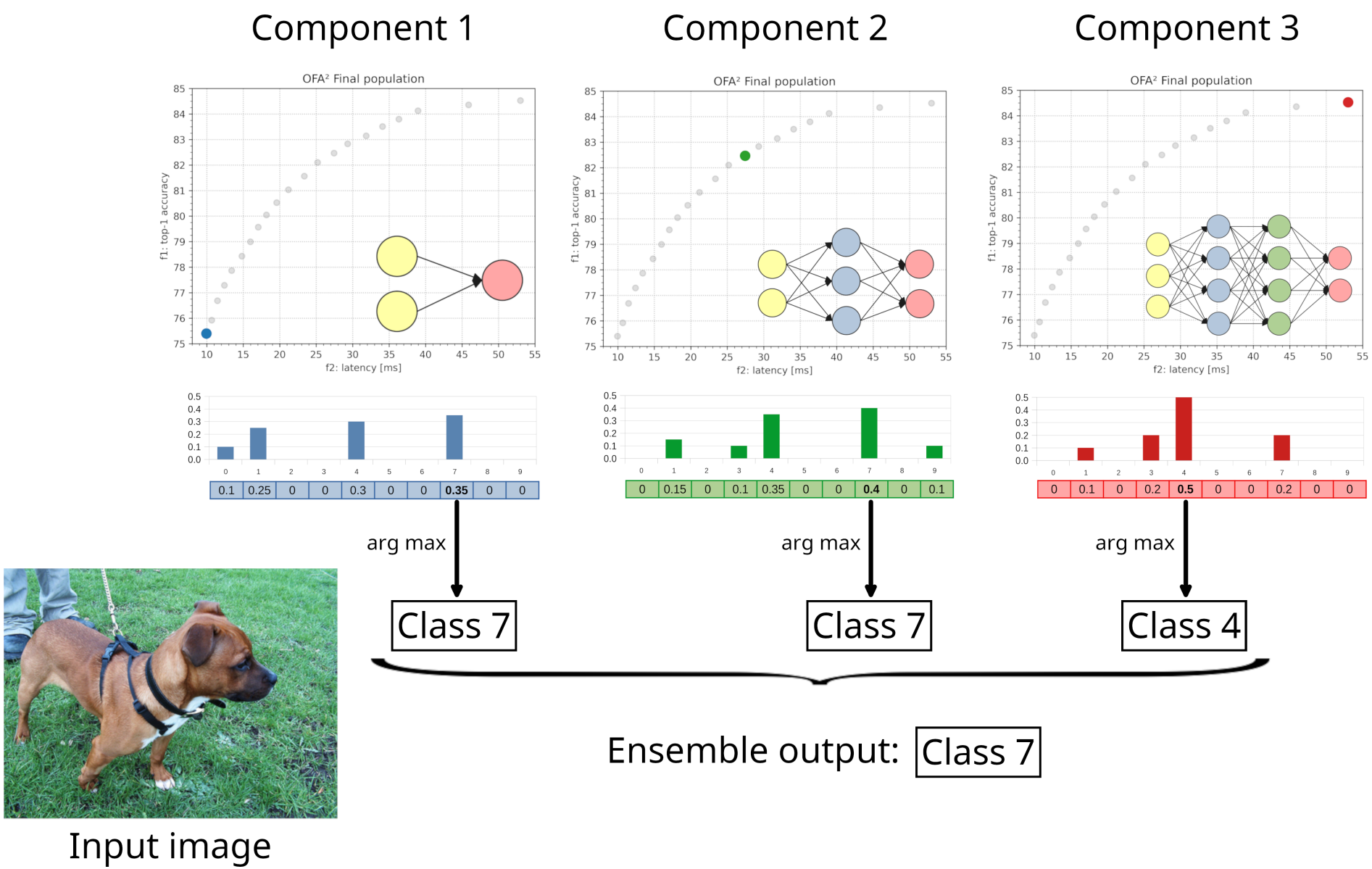
Overview

- **Voting**:
  - Hard voting
  - Soft voting
- **Latency**:
  - Sum
  - Maximum
- **Formation**:
  - Manual Sampling
  - Multi-Objective Optimization

Voting

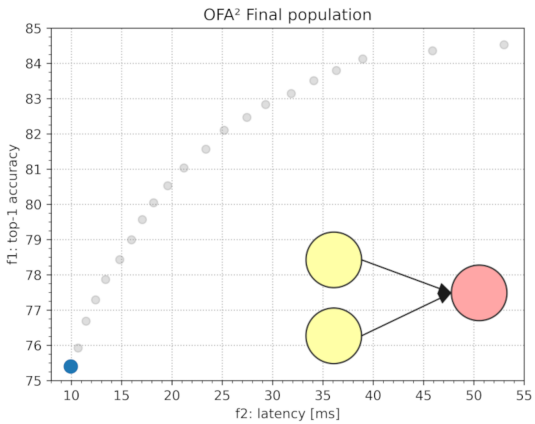
- **Voting**:
  - Hard voting
  - Soft voting

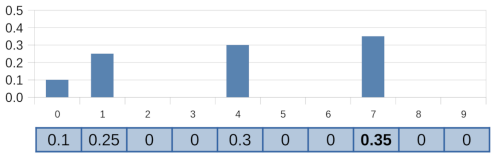
Hard voting


Component 1
Component 2
Component 3
OFA² Final population
f1: top-1 accuracy
f2: latency [ms]
0.1 0.25 0 0 0.3 0 0 0.35 0 0
0 0.15 0 0.1 0.35 0 0 0.4 0 0.1
0 0.1 0 0.2 0.5 0 0 0.2 0 0
arg max
arg max
arg max
Class 7
Class 7
Class 4
Ensemble output: Class 7
Input image

Soft voting

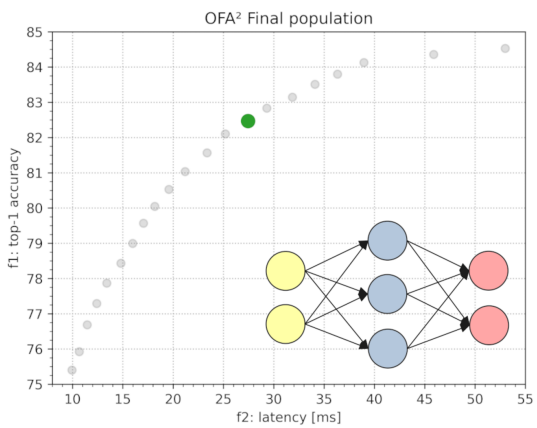
Component 1

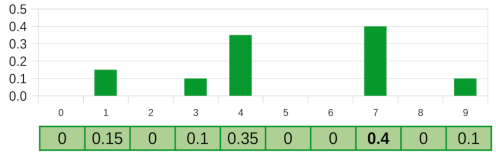




0	1	2	3	4	5	6	7	8	9
0.1	0.25	0	0	0.3	0	0	**0.35**	0	0

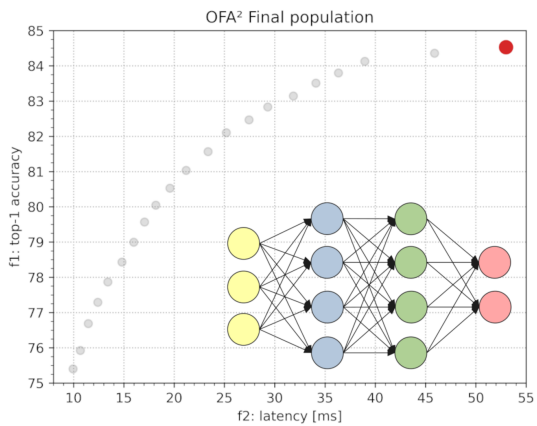
Component 2

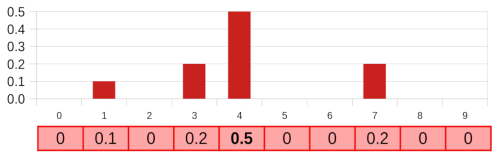




0	1	2	3	4	5	6	7	8	9
0	0.15	0	0.1	0.35	0	0	**0.4**	0	0.1

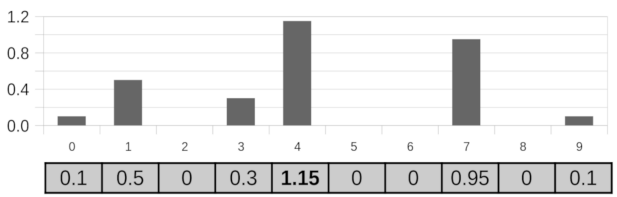
Component 3





0	1	2	3	4	5	6	7	8	9
0	0.1	0	0.2	**0.5**	0	0	0.2	0	0

Input image



0	1	2	3	4	5	6	7	8	9
0.1	0.5	0	0.3	**1.15**	0	0	0.95	0	0.1

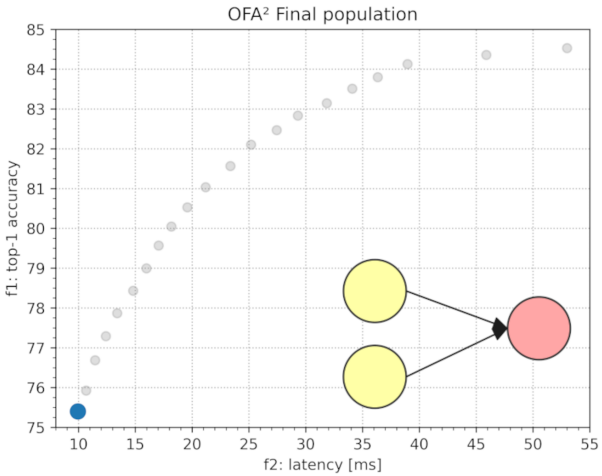
Ensemble output: arg max → Class 4

Latency

- **Latency**:
  - Sum
  - Maximum

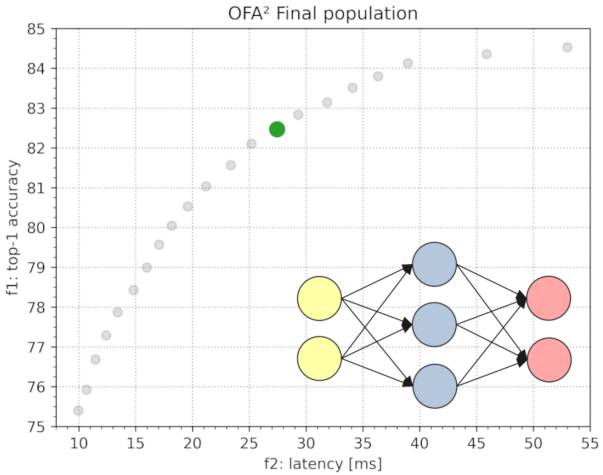
Sum

Component 1



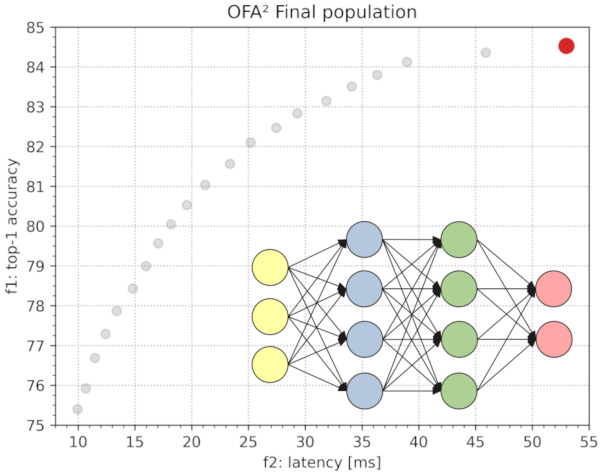
Latency = 10 ms

Component 2



Latency = 30 ms

Component 3



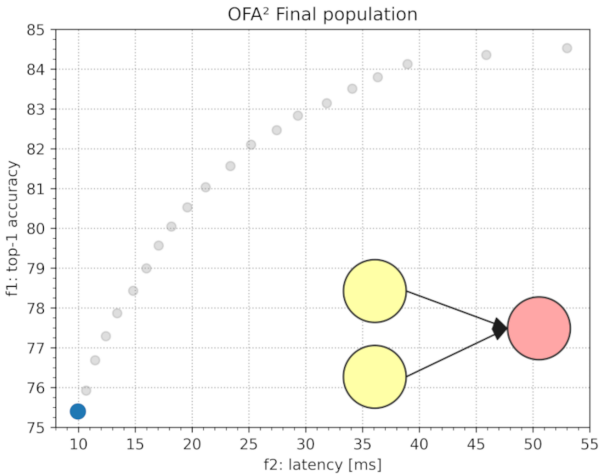
Latency = 50 ms

+

Ensemble latency = 90 ms

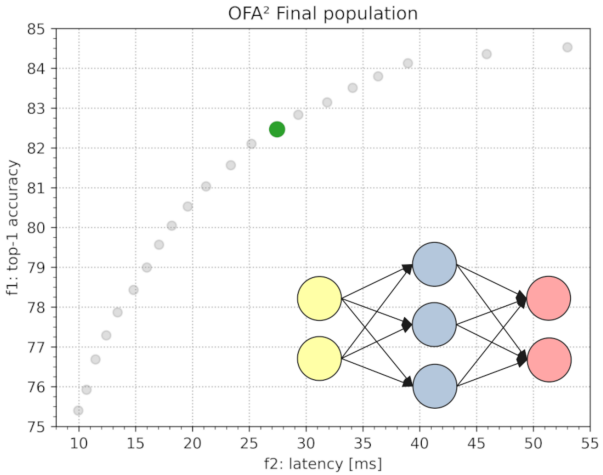
Maximum

Component 1



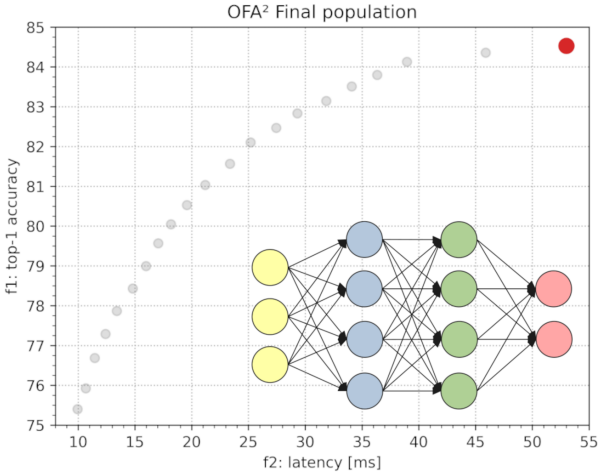
Latency = 10 ms

Component 2



Latency = 30 ms

Component 3



Latency = 50 ms

MAX

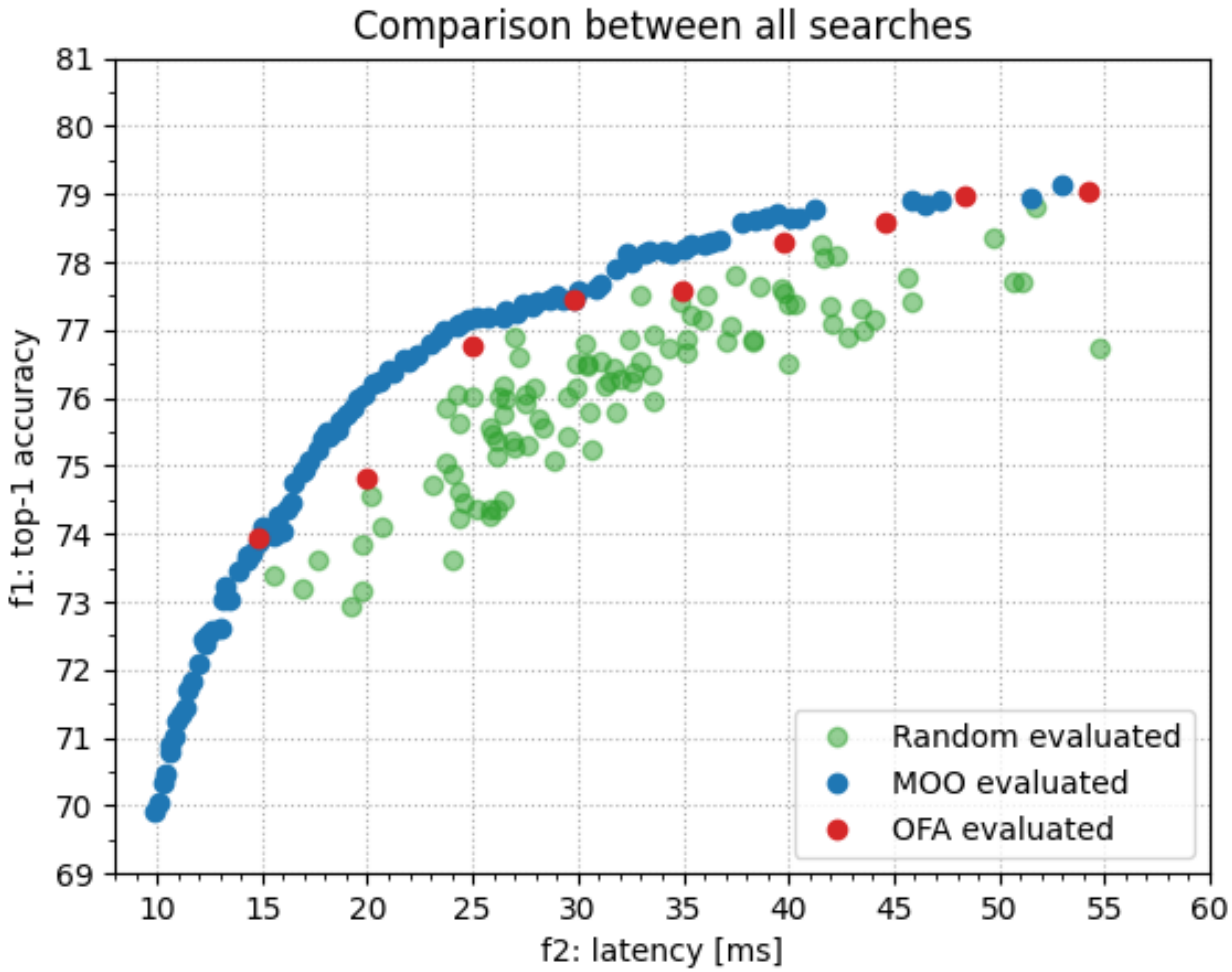
Ensemble latency = 50 ms

Agenda

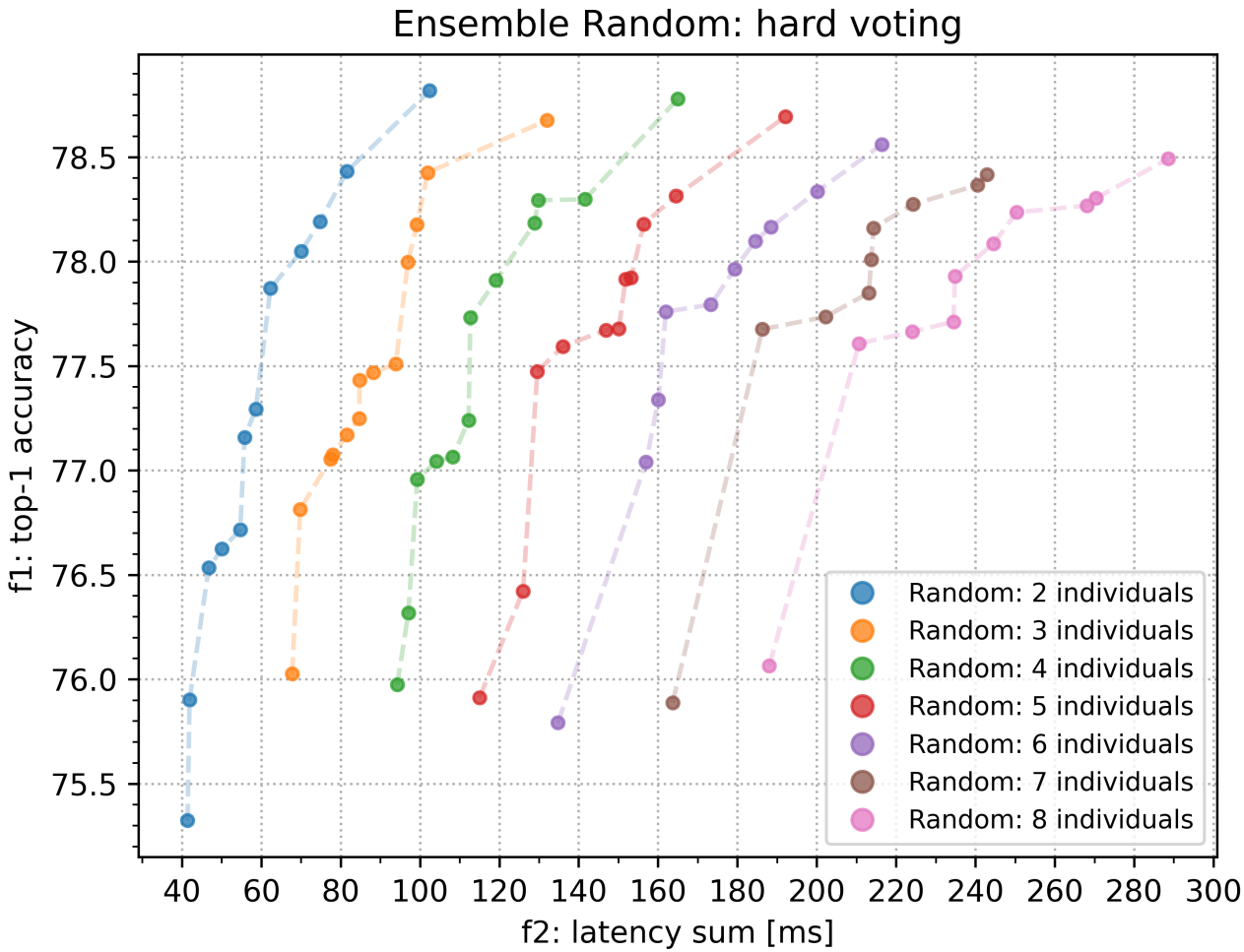
Overview

- **Experiments**:
  - 43 ensembles for each scenario
- **Individuals**:
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals
- **Voting**:
  - Hard voting
  - Soft voting
- **Latency**:
  - Sum
  - Maximum

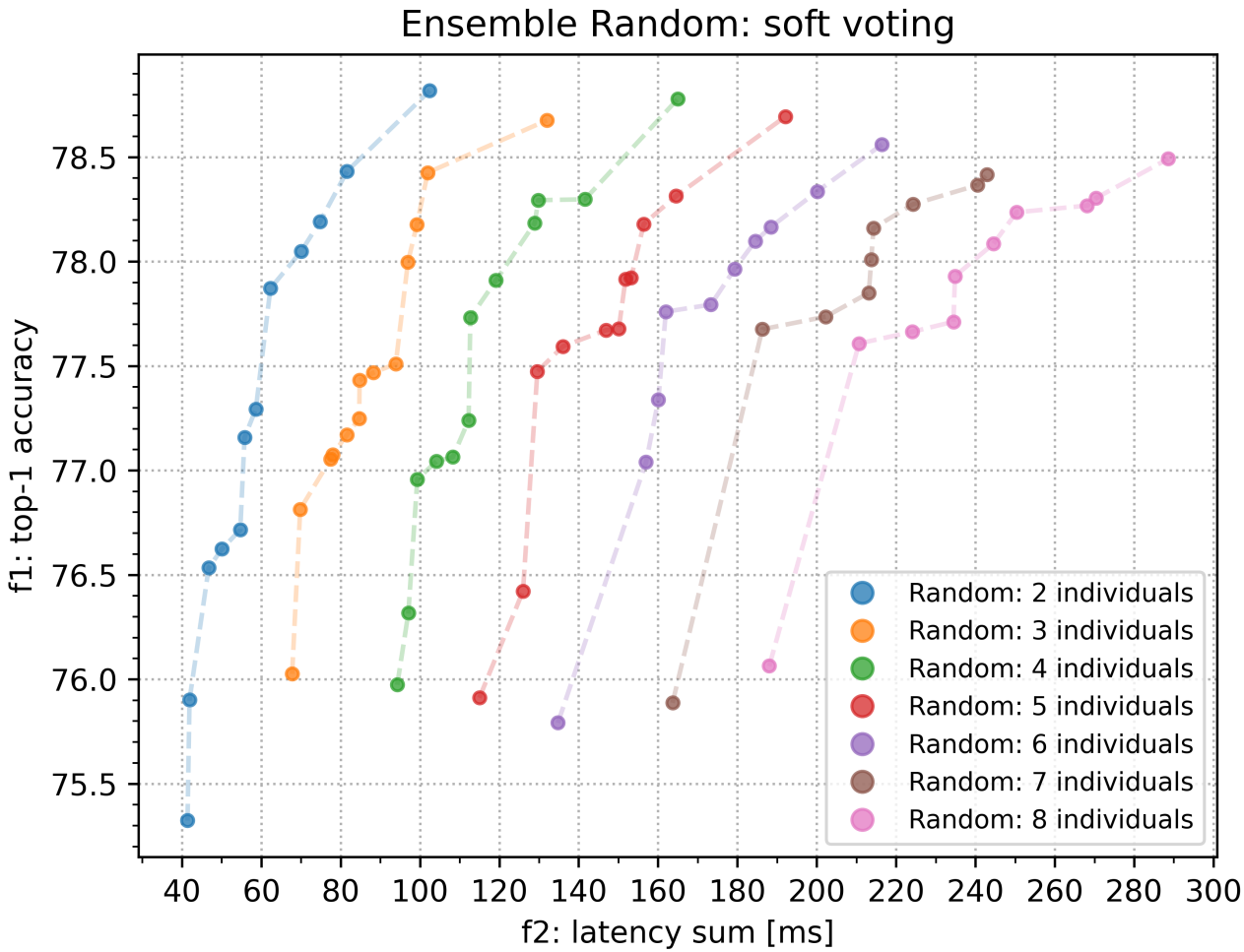
Random search



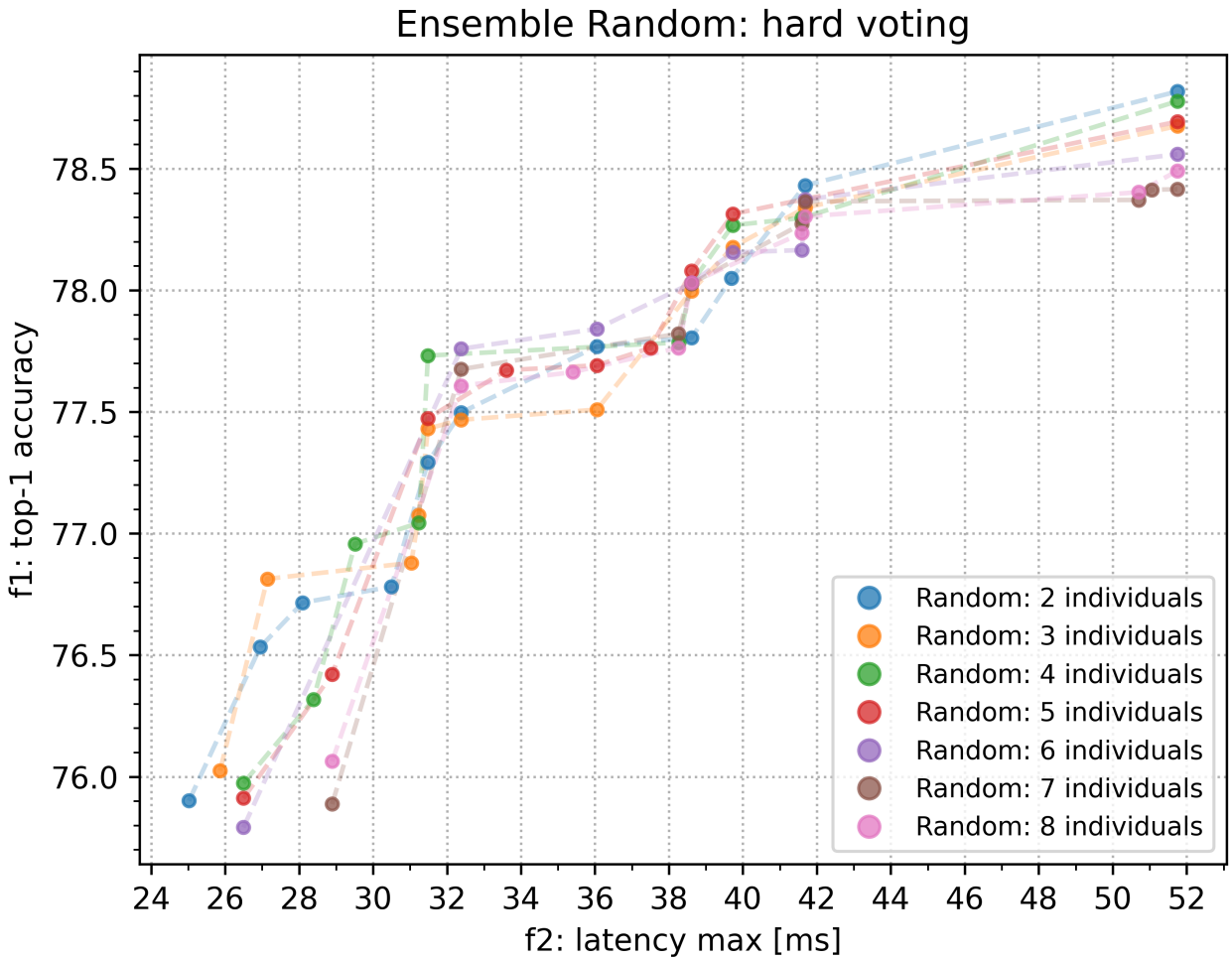
Random search: hard, sum



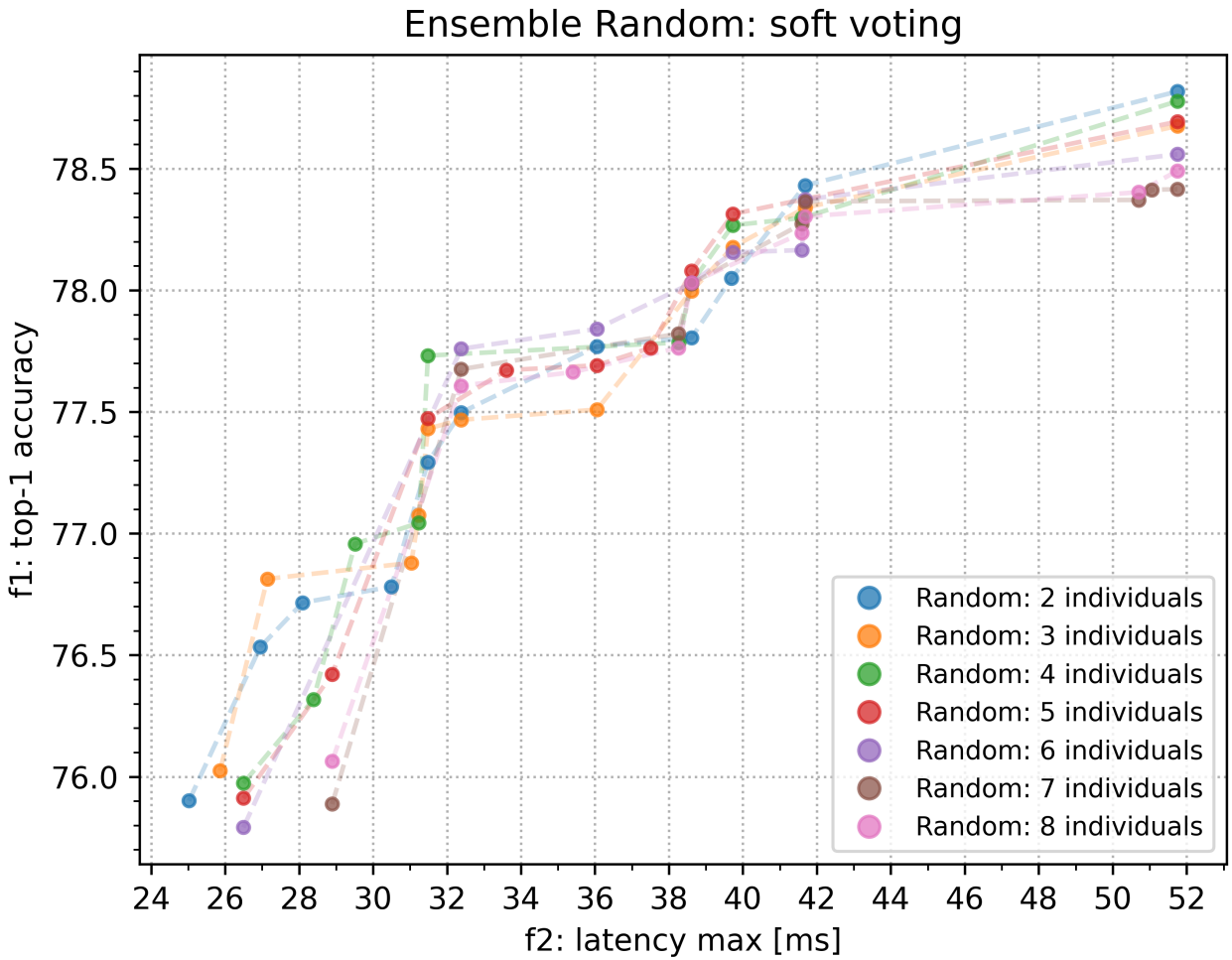
Random search: soft, sum



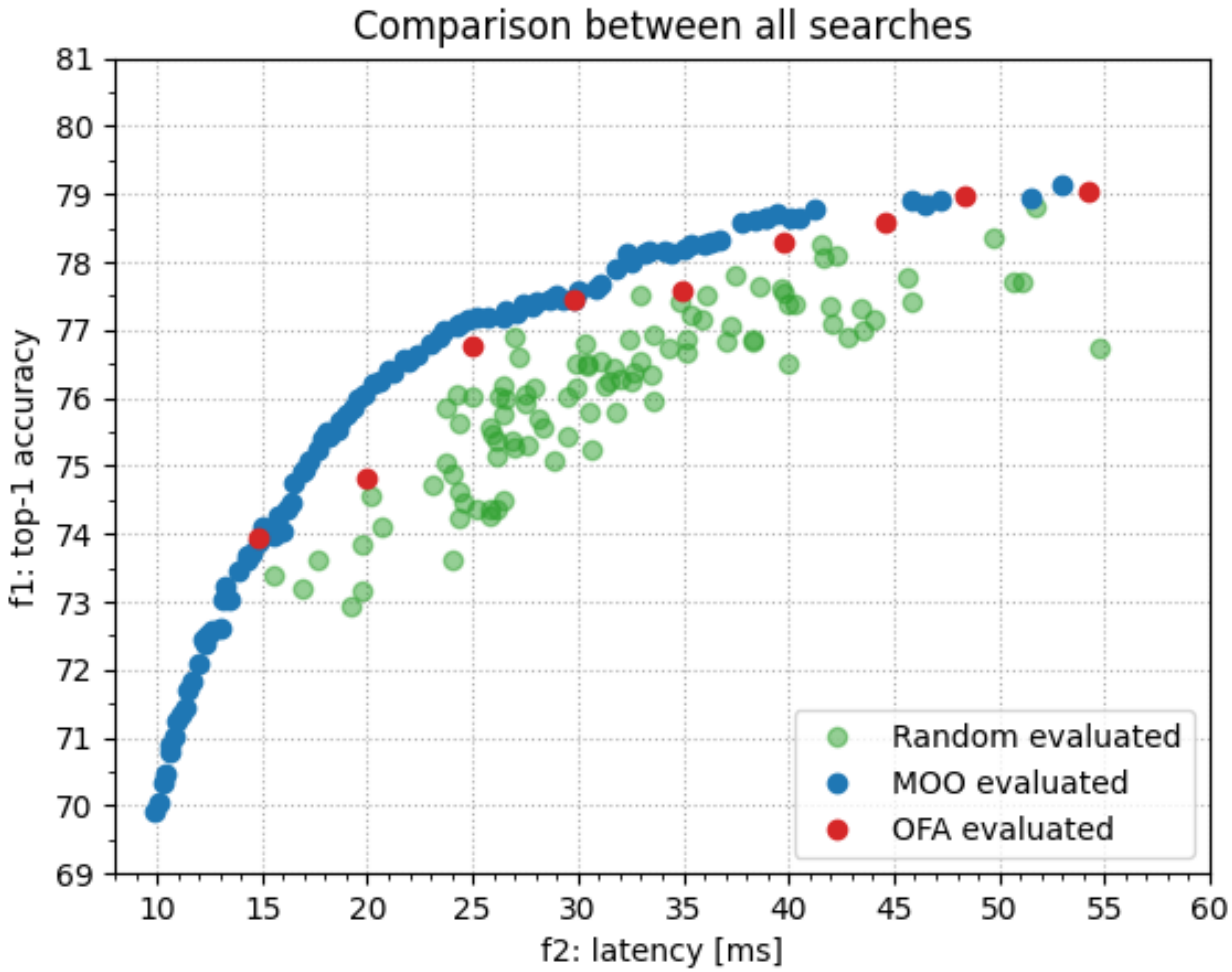
Random search: hard, max



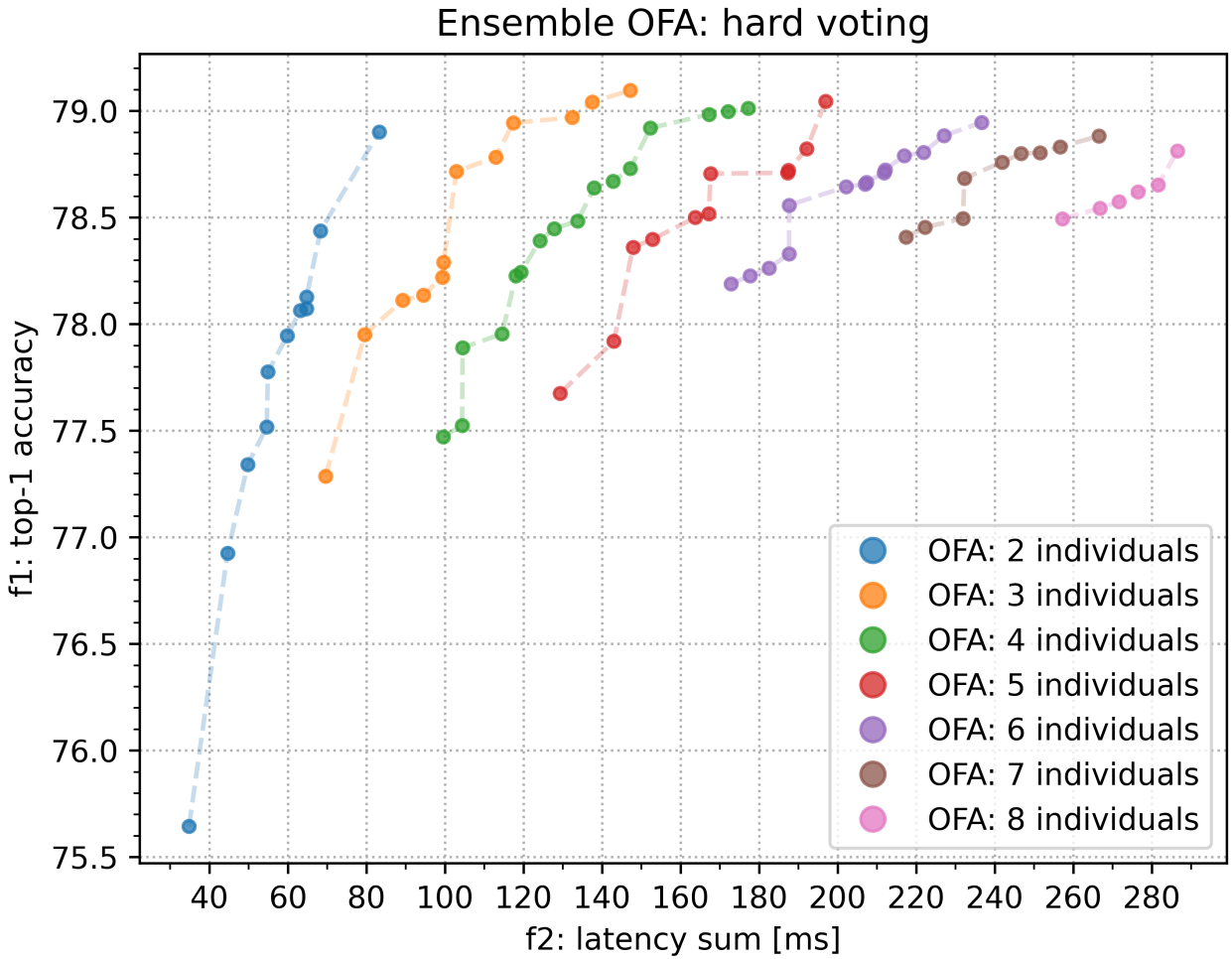
Random search: soft, max



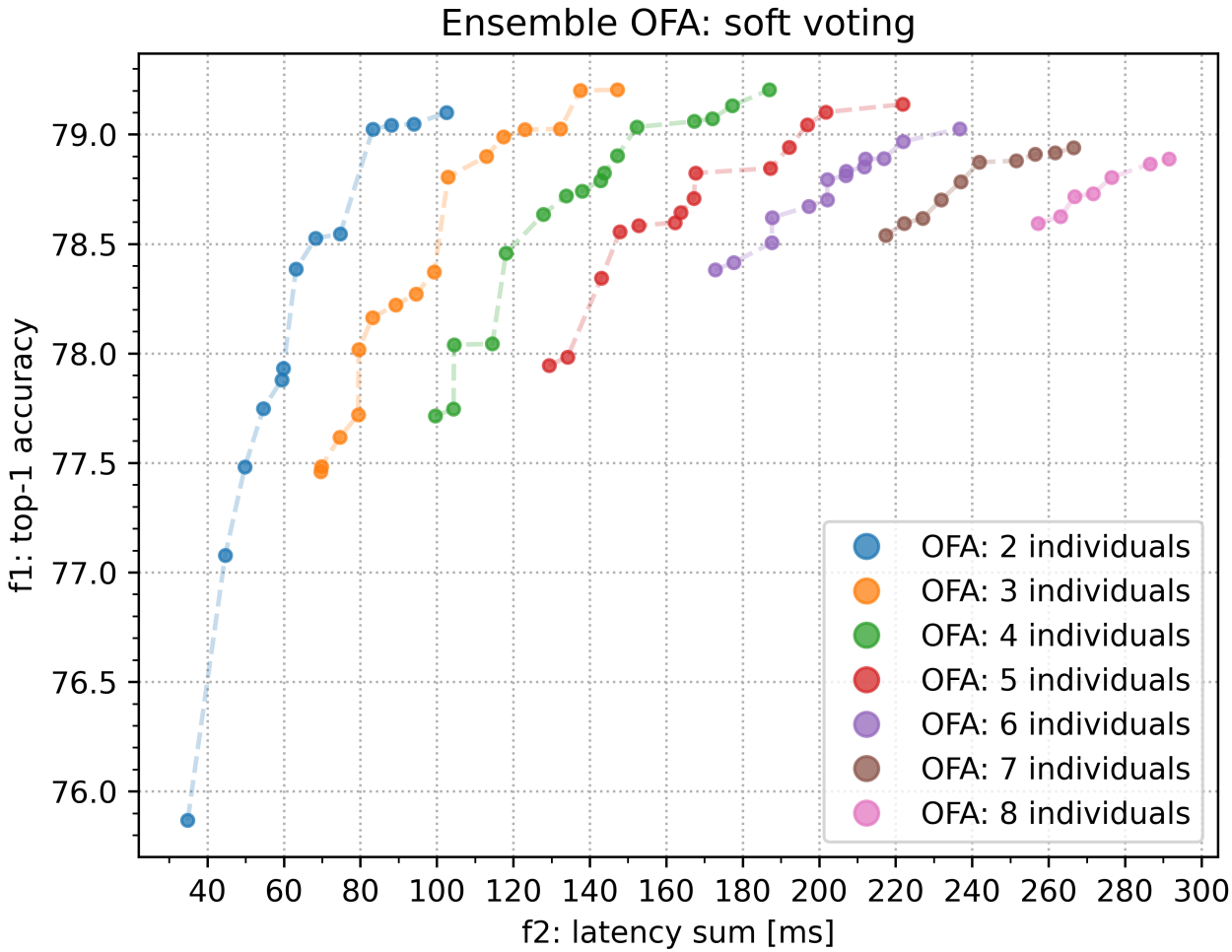
OFA search


Comparison between all searches
f1: top-1 accuracy
f2: latency [ms]
Random evaluated
MOO evaluated
OFA evaluated

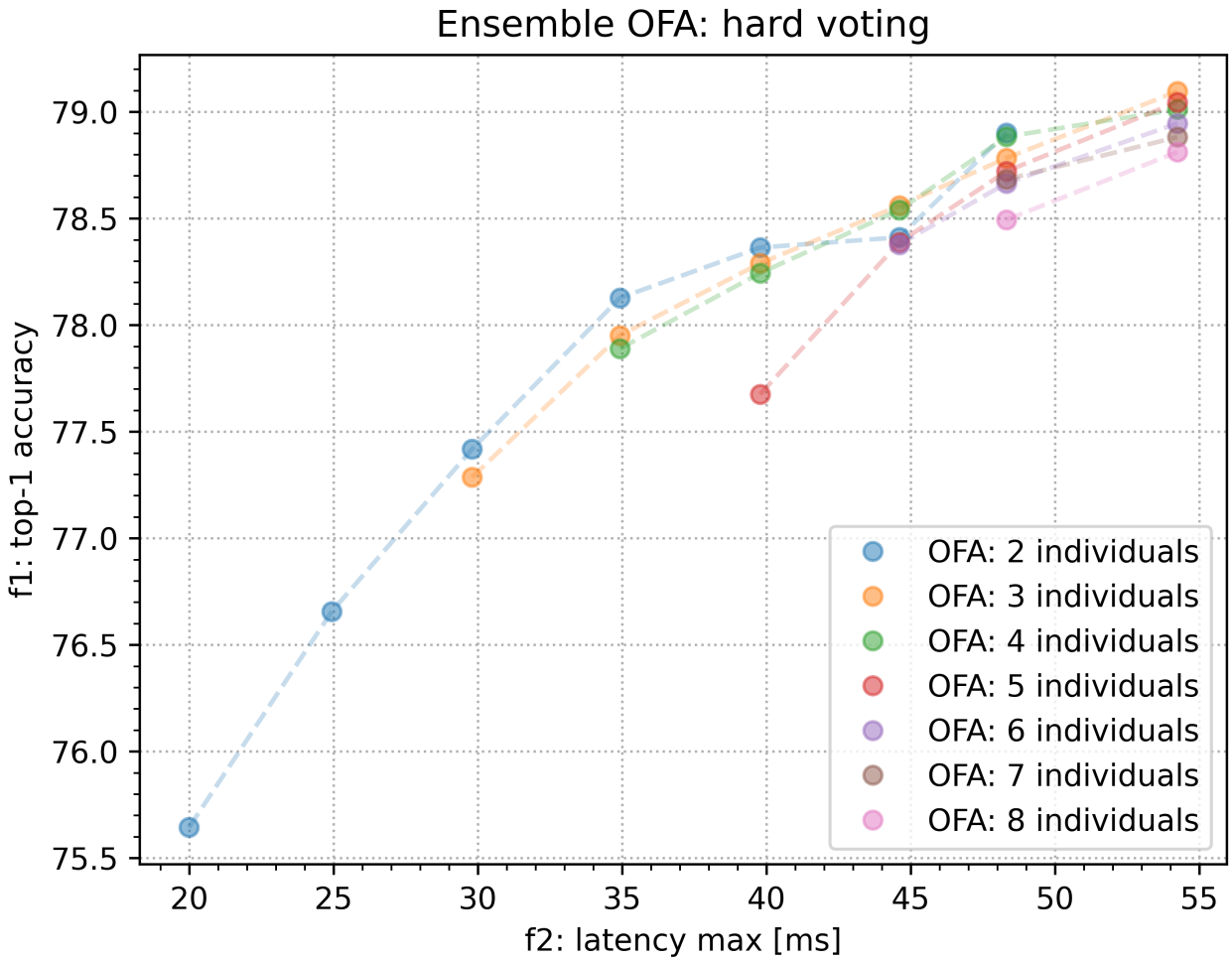
OFA search: hard, sum



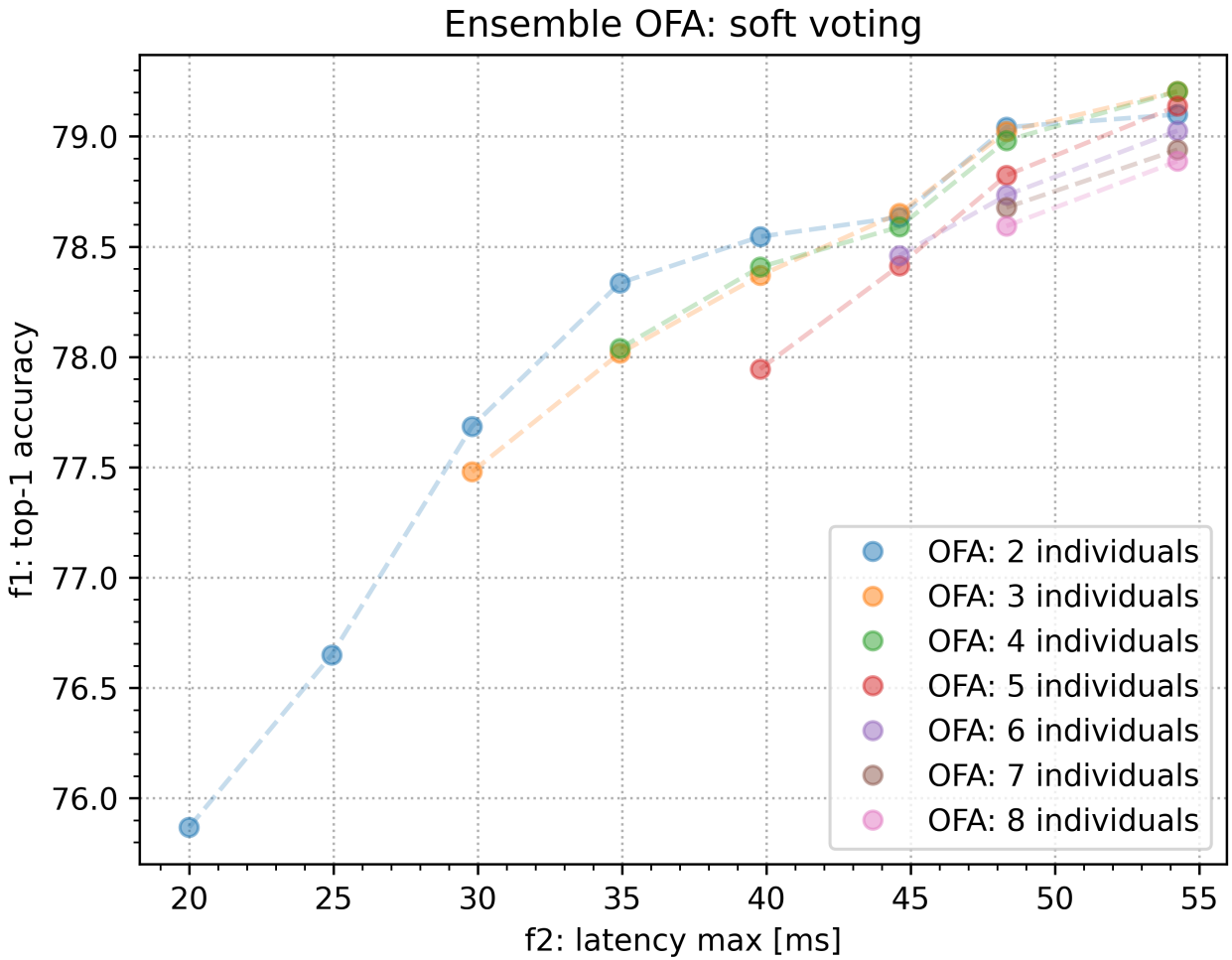
OFA search: soft, sum



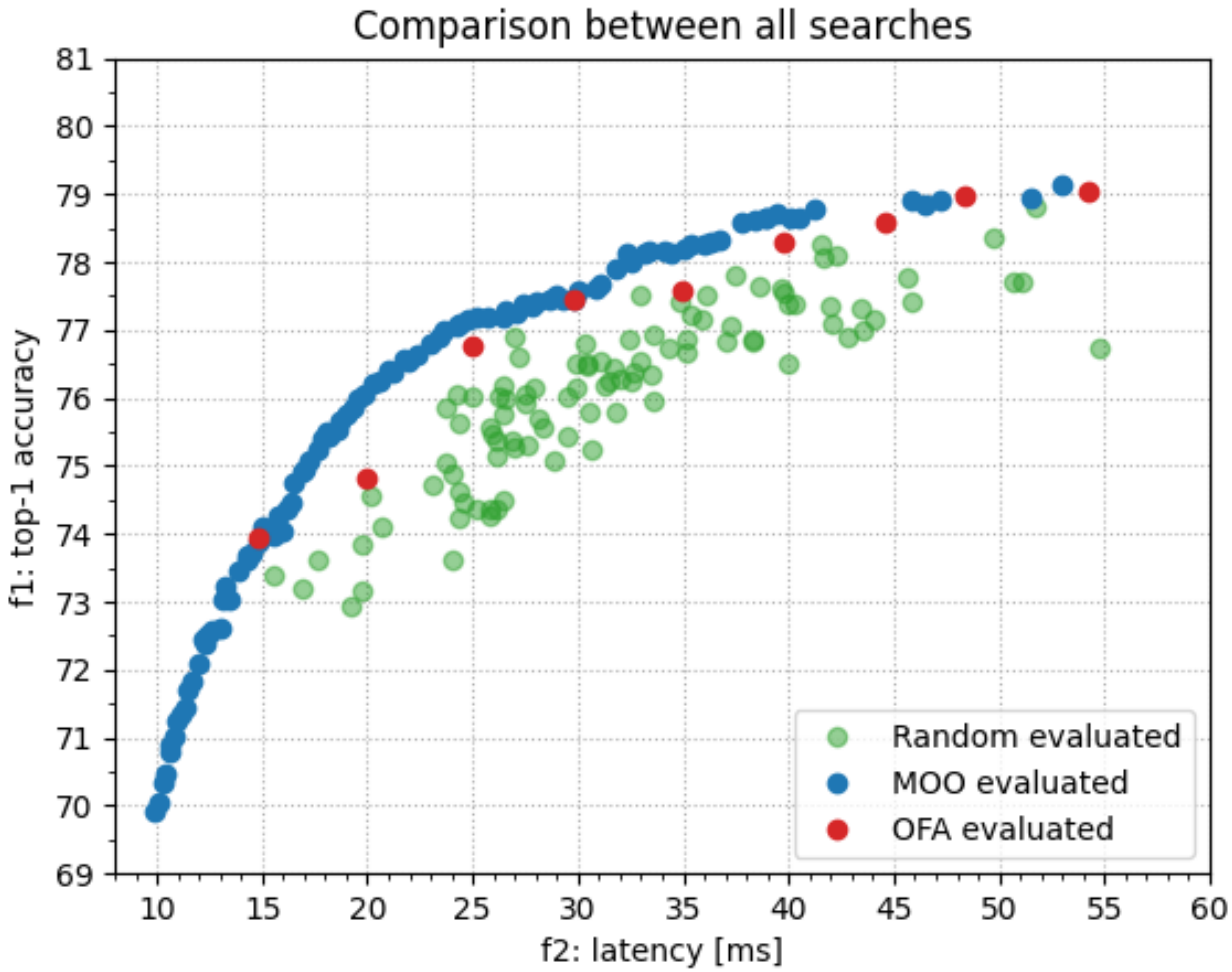
OFA search: hard, max



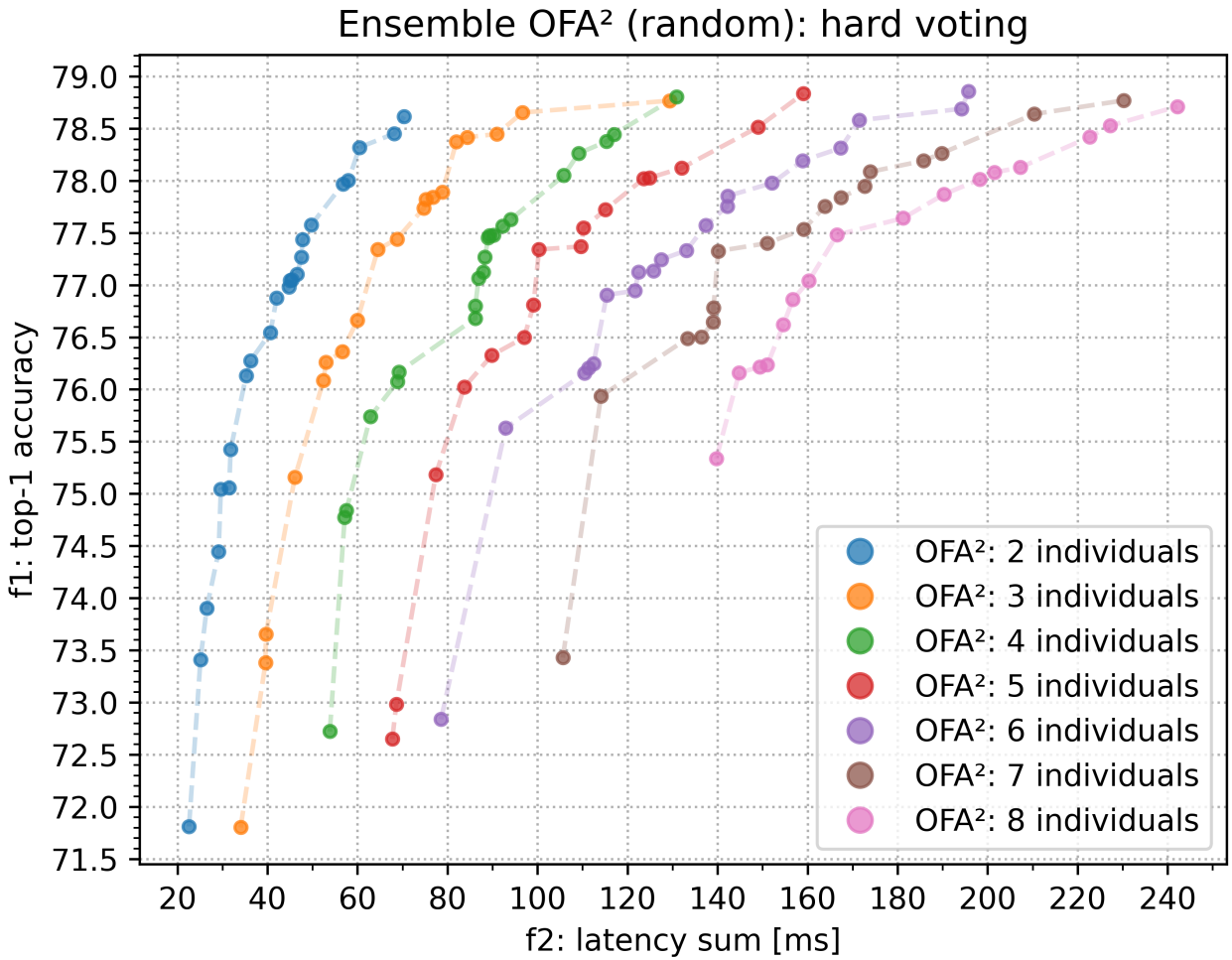
OFA search: soft, max



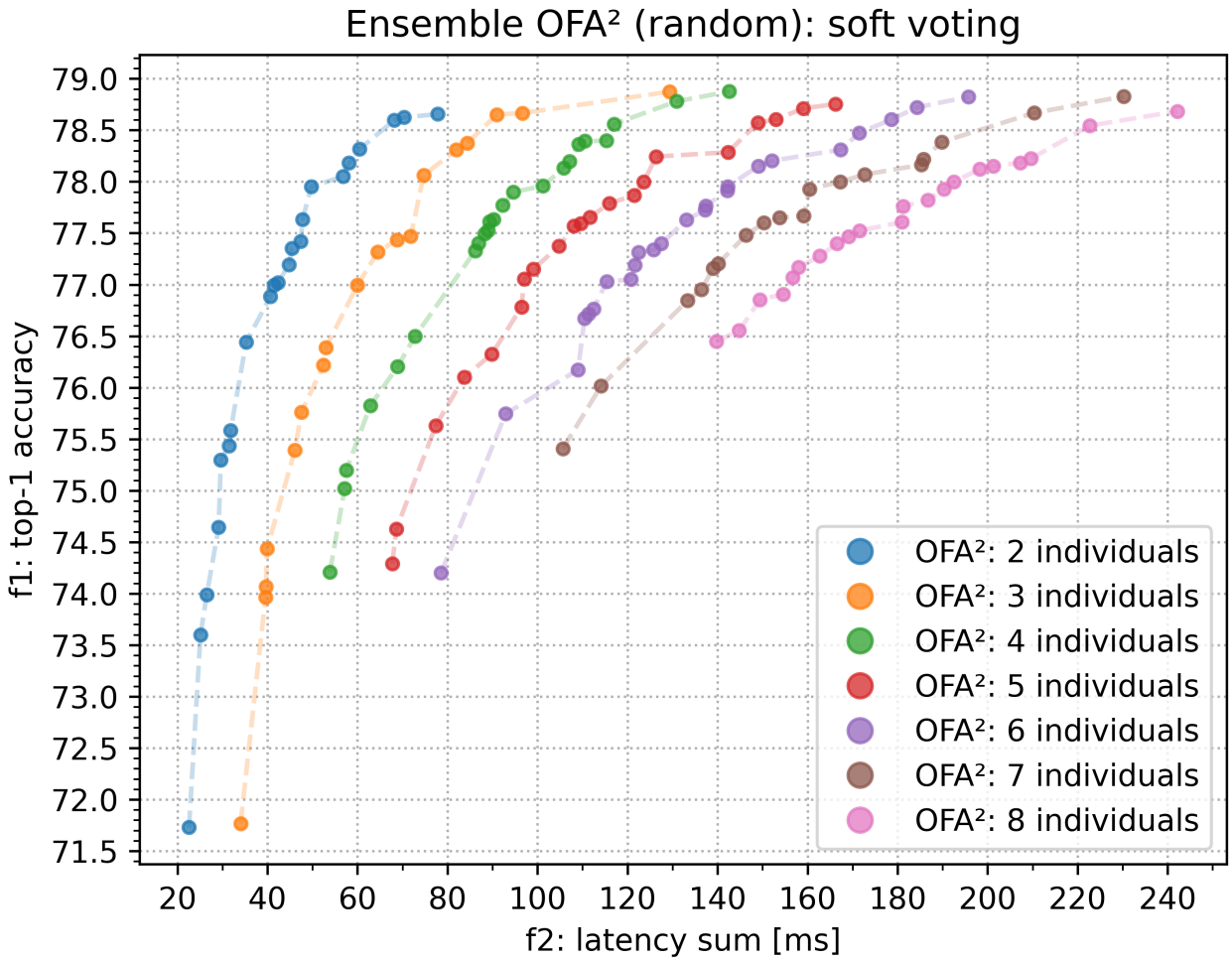
OFA² search (random)



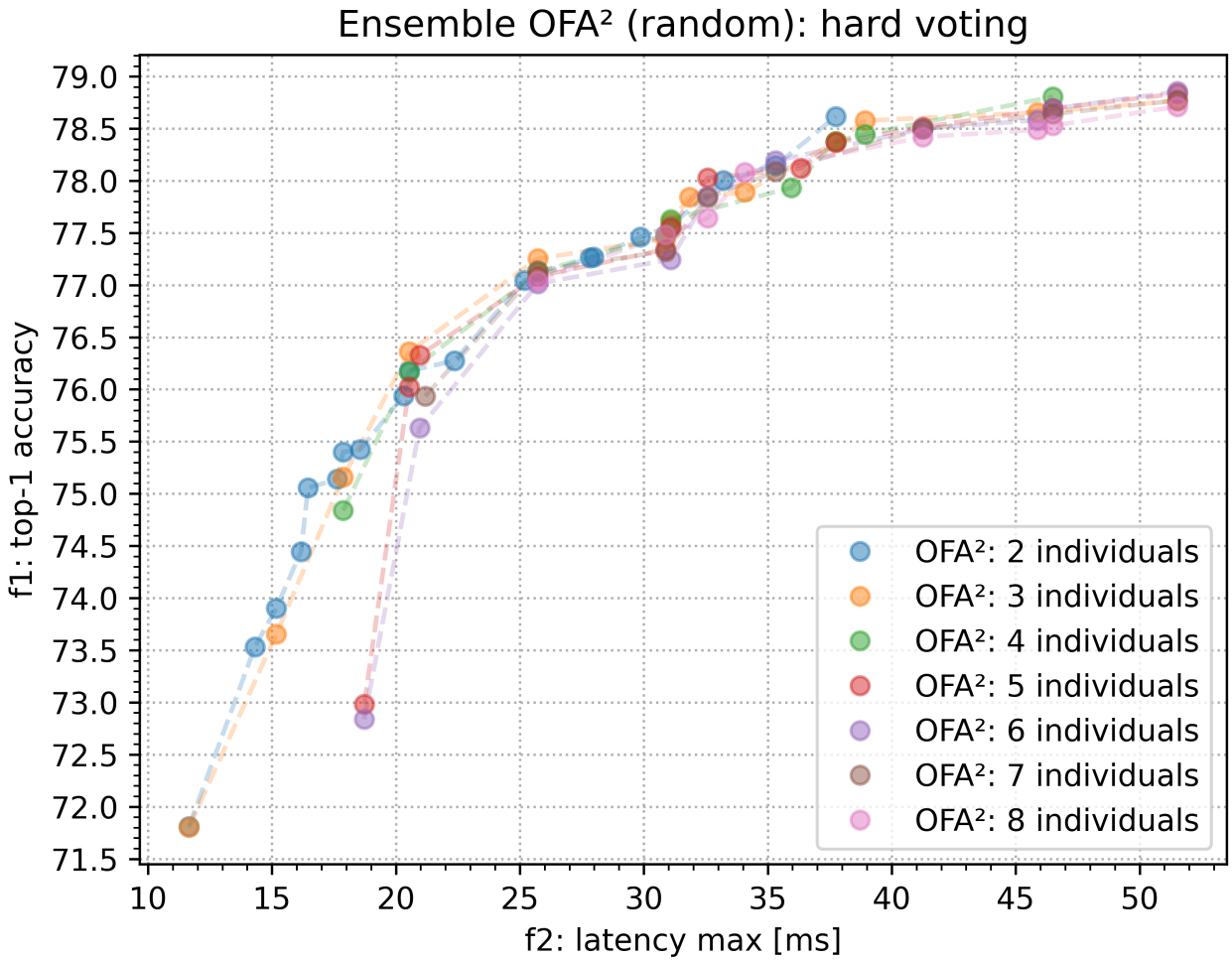
OFA² search (random): hard, sum



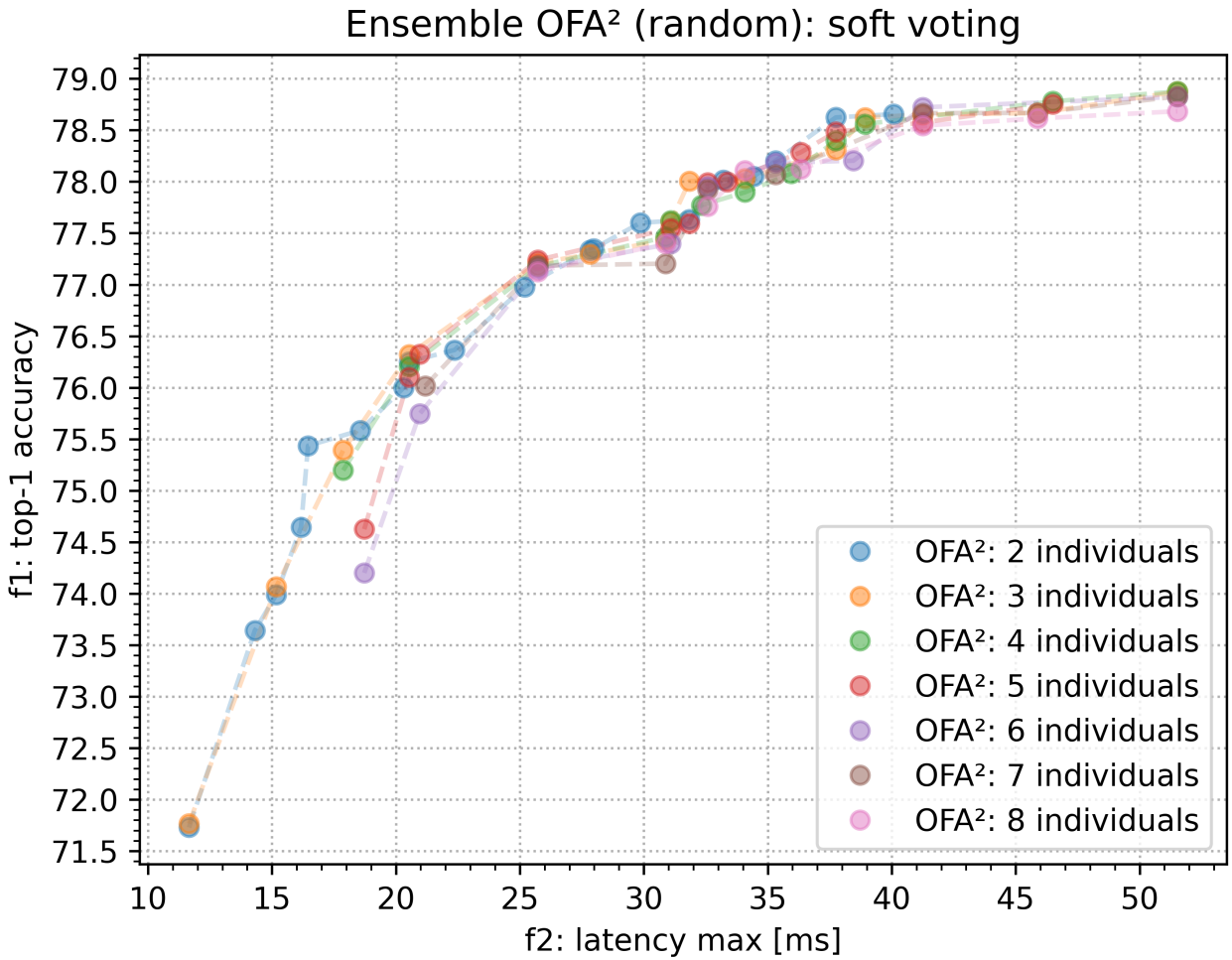
OFA² search (random): soft, sum



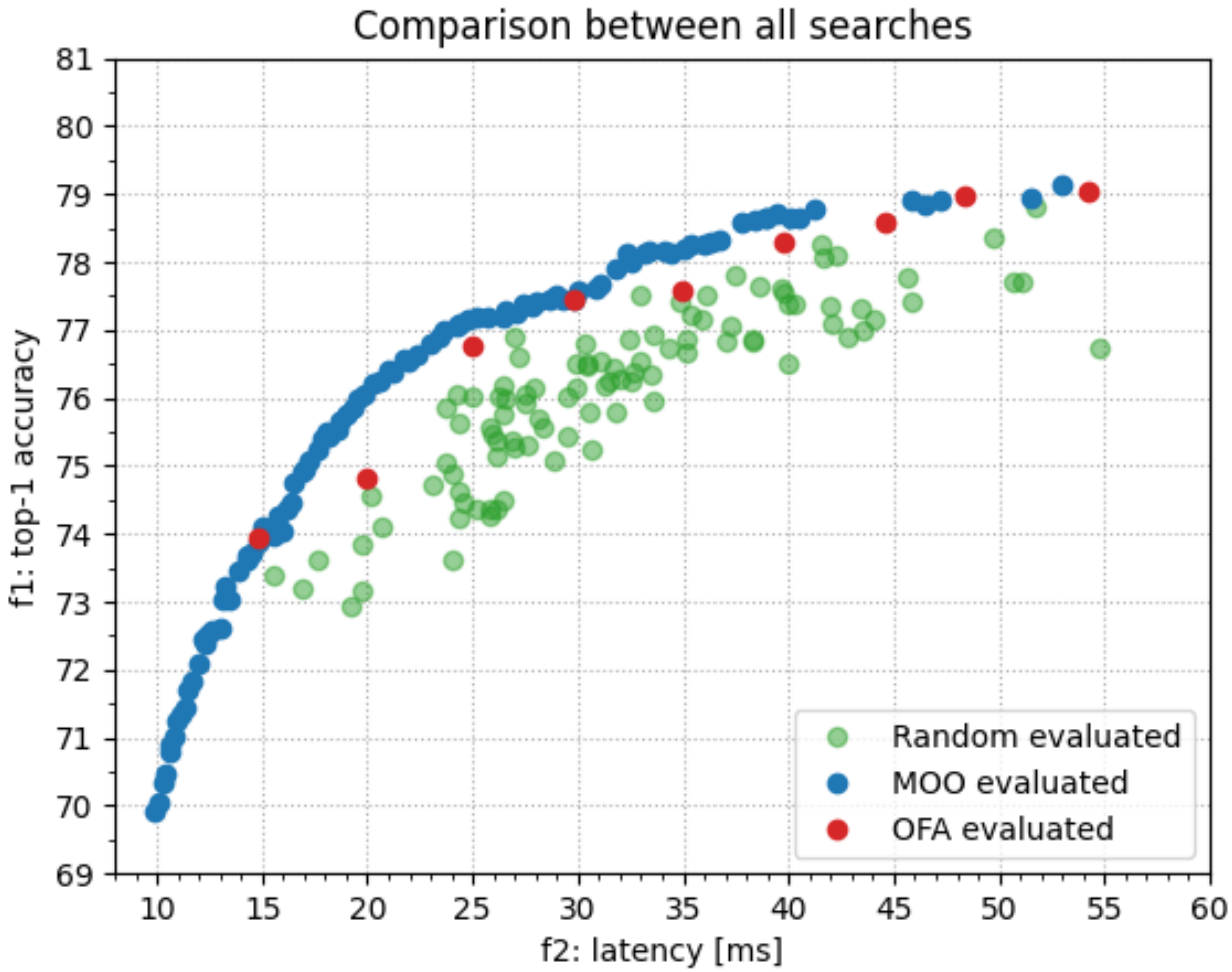
OFA² search (random): hard, max



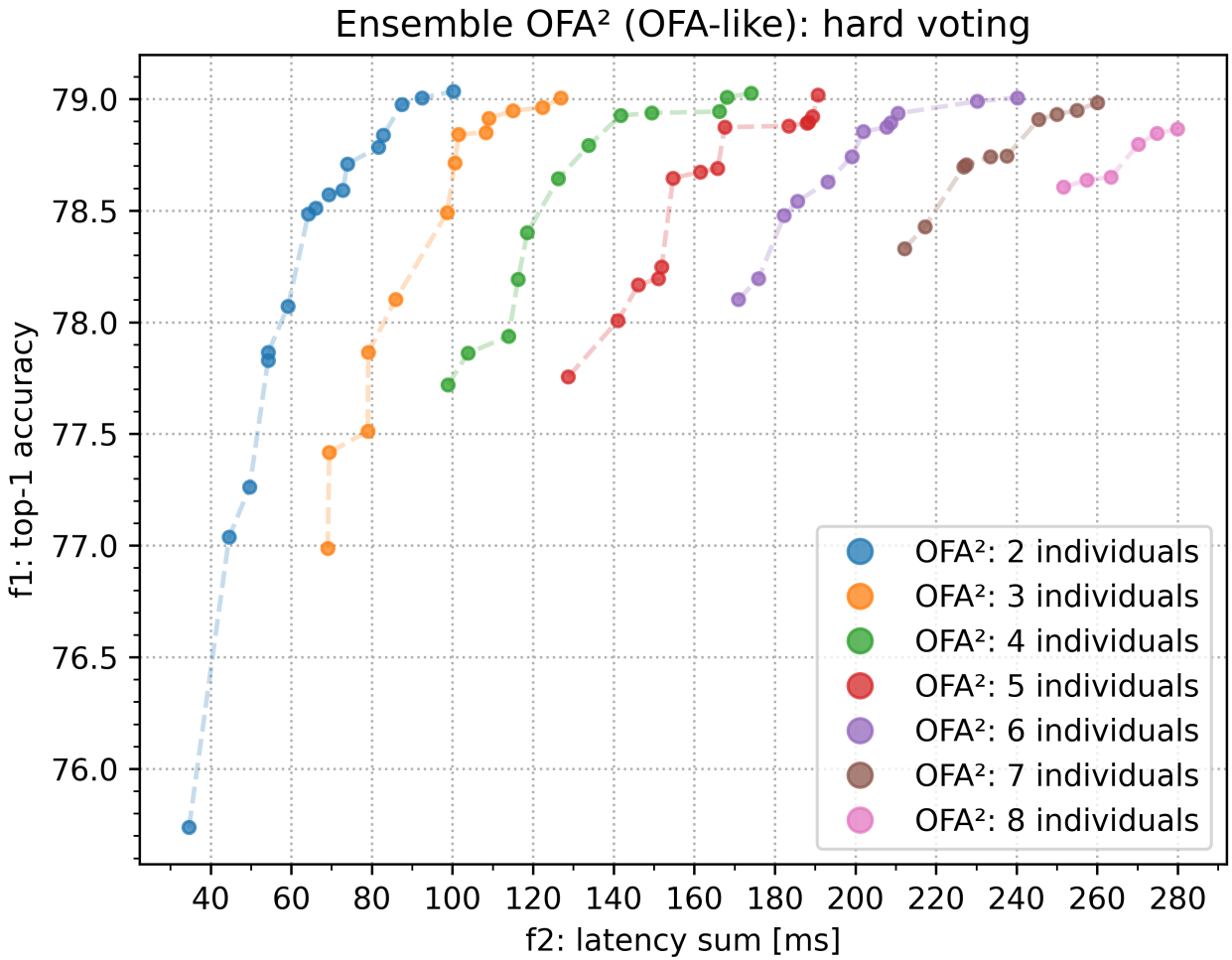
OFA² search (random): soft, sum



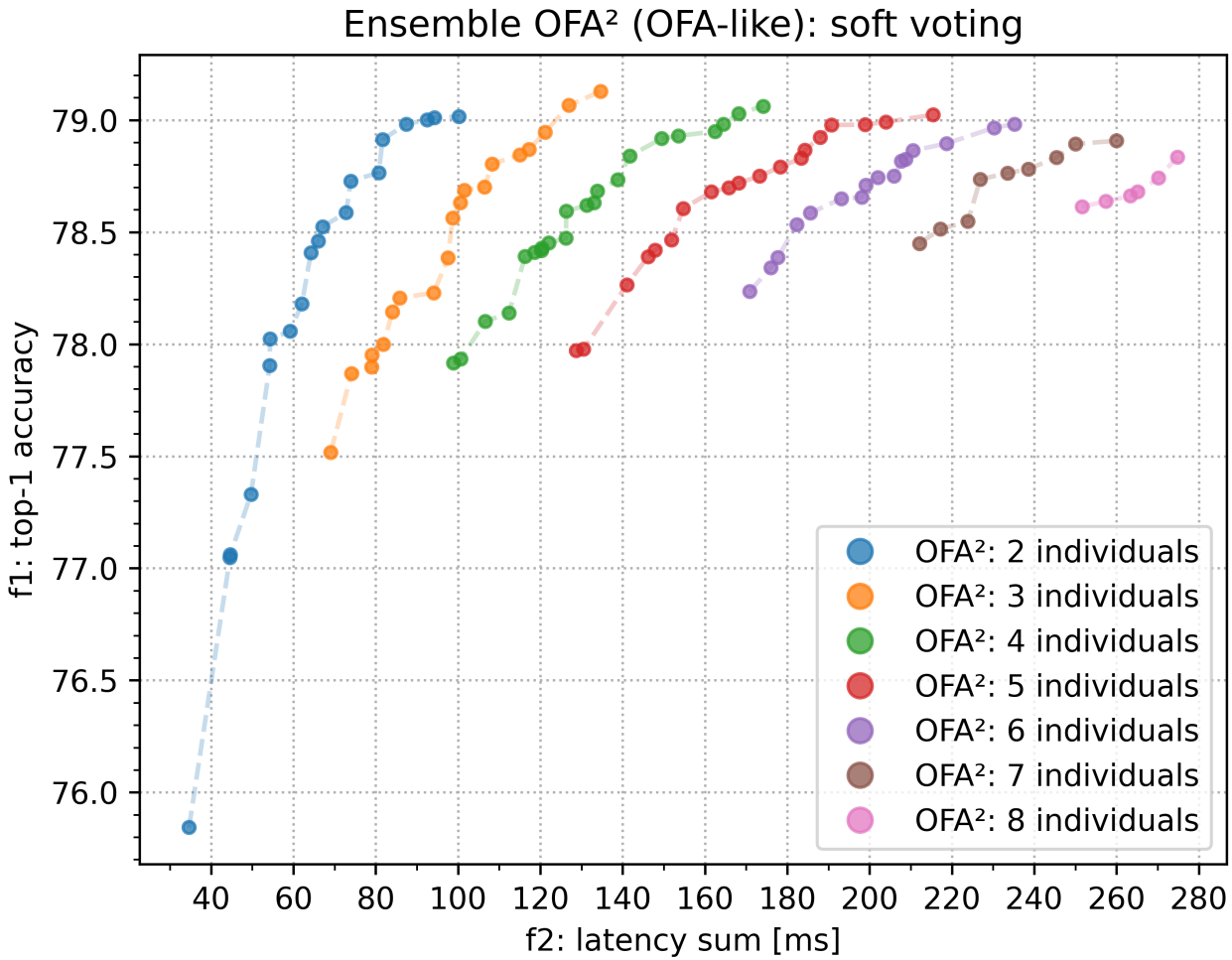
OFA² search (OFA-like)



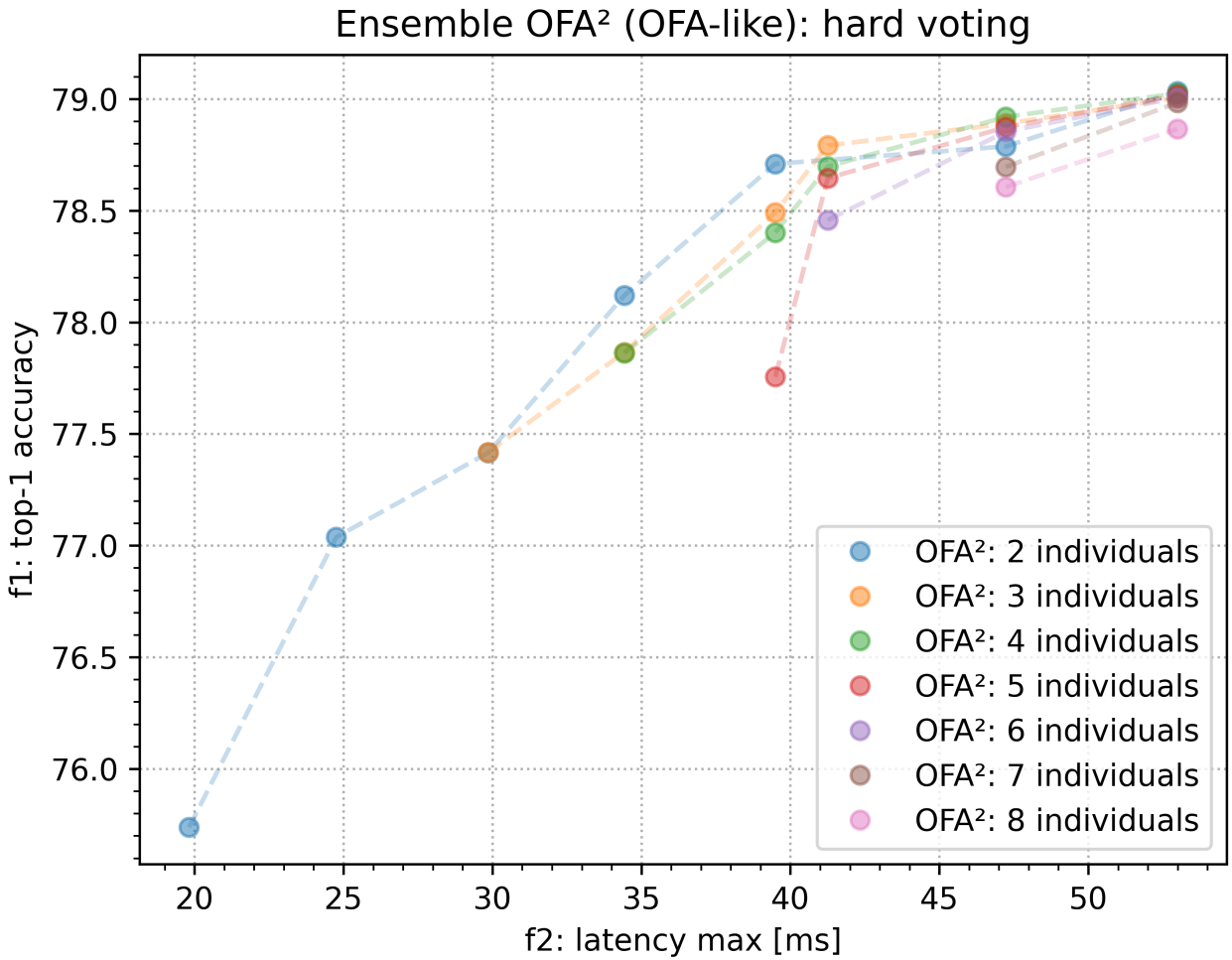
OFA² search (OFA-like): hard, sum



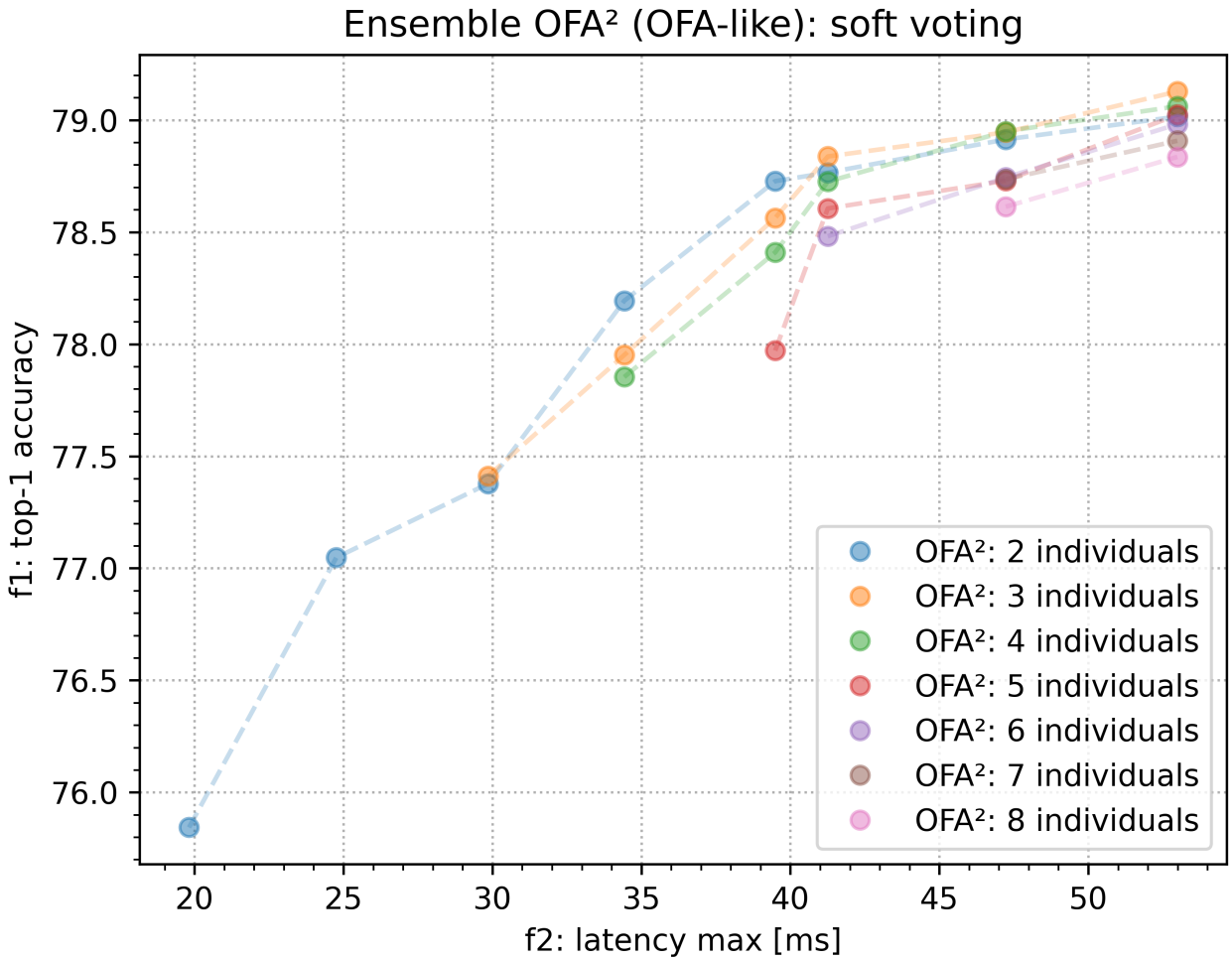
OFA² search (OFA-like): soft, sum



OFA² search (OFA-like): hard, max



OFA² search (OFA-like): soft, sum

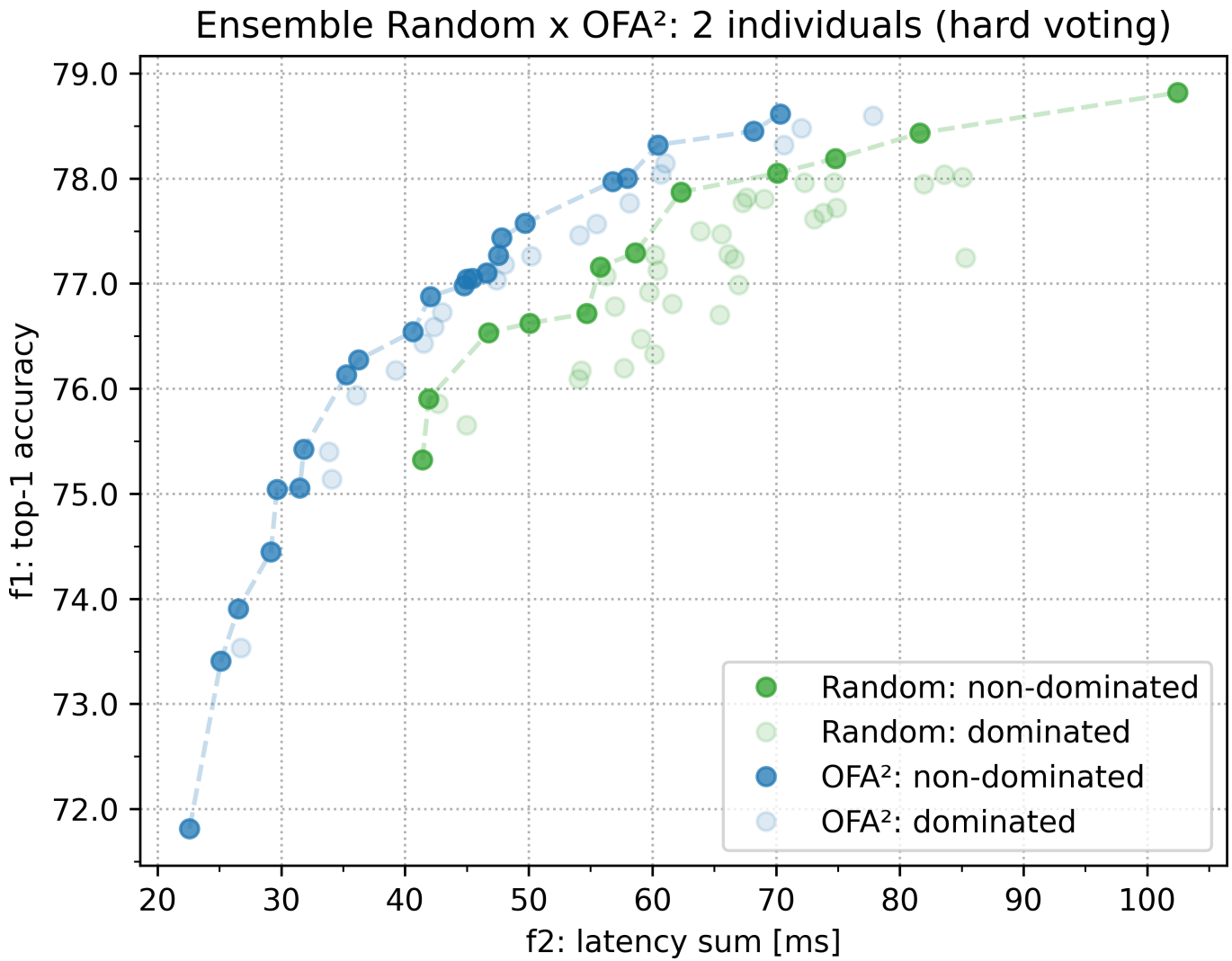


OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

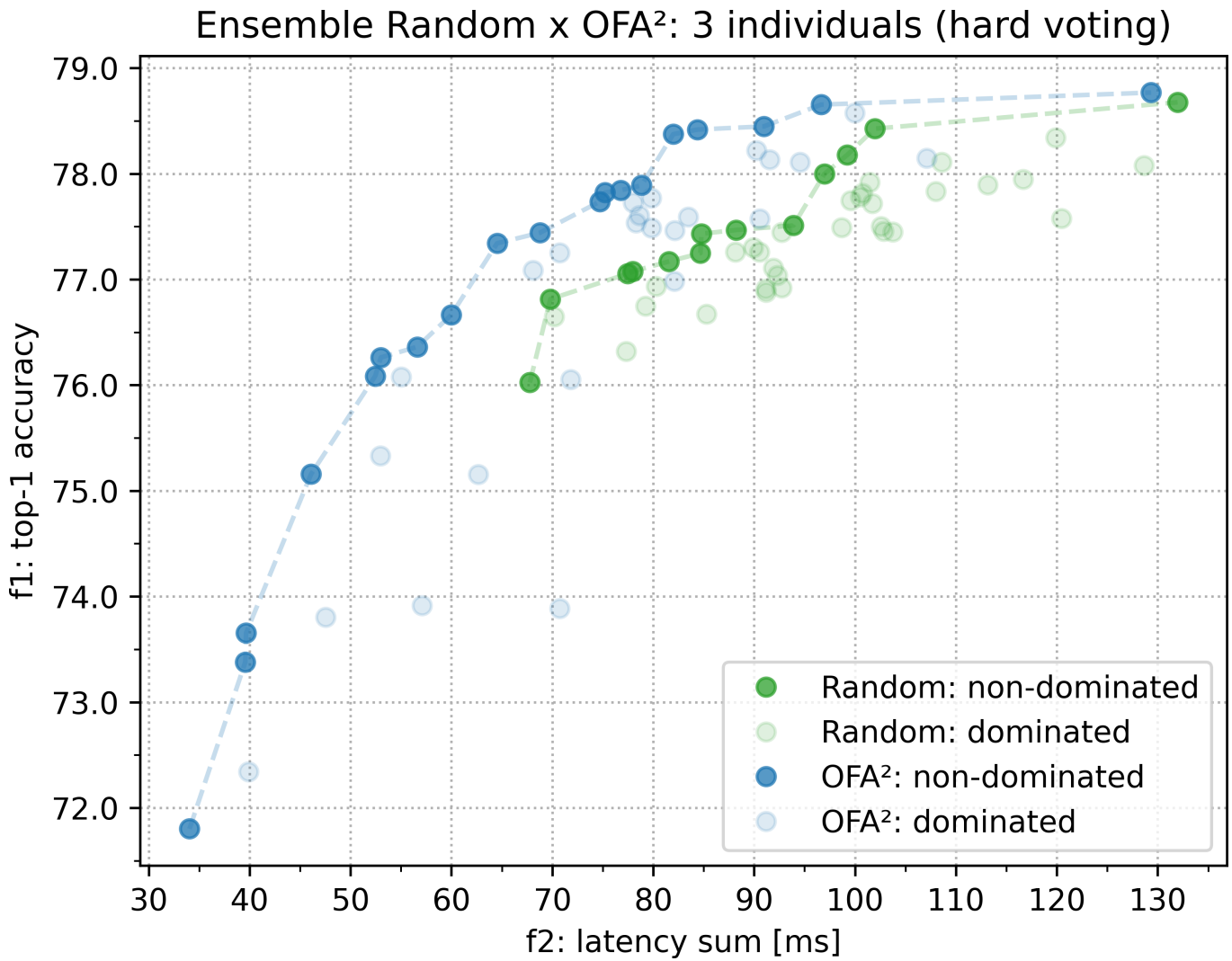
OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



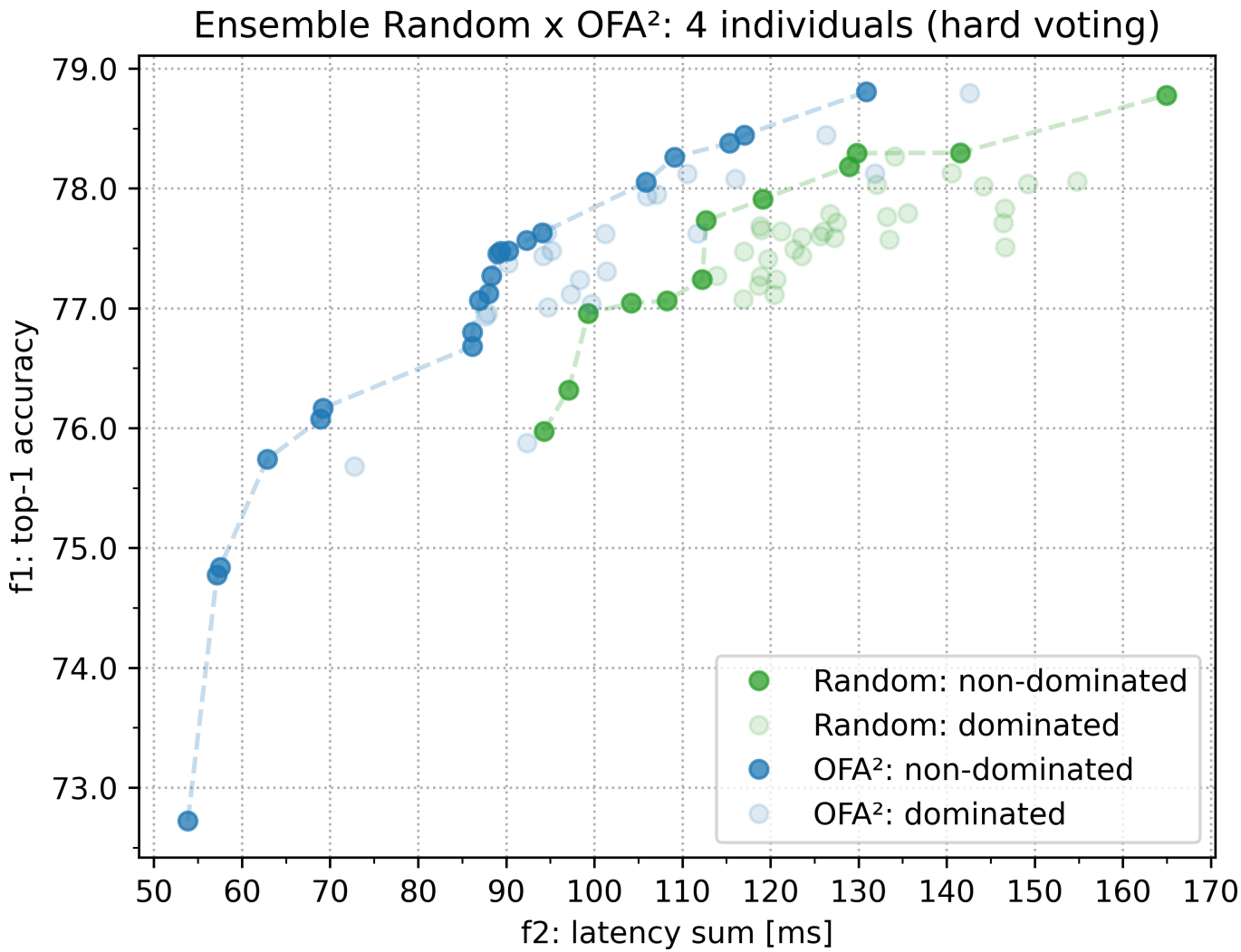
OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



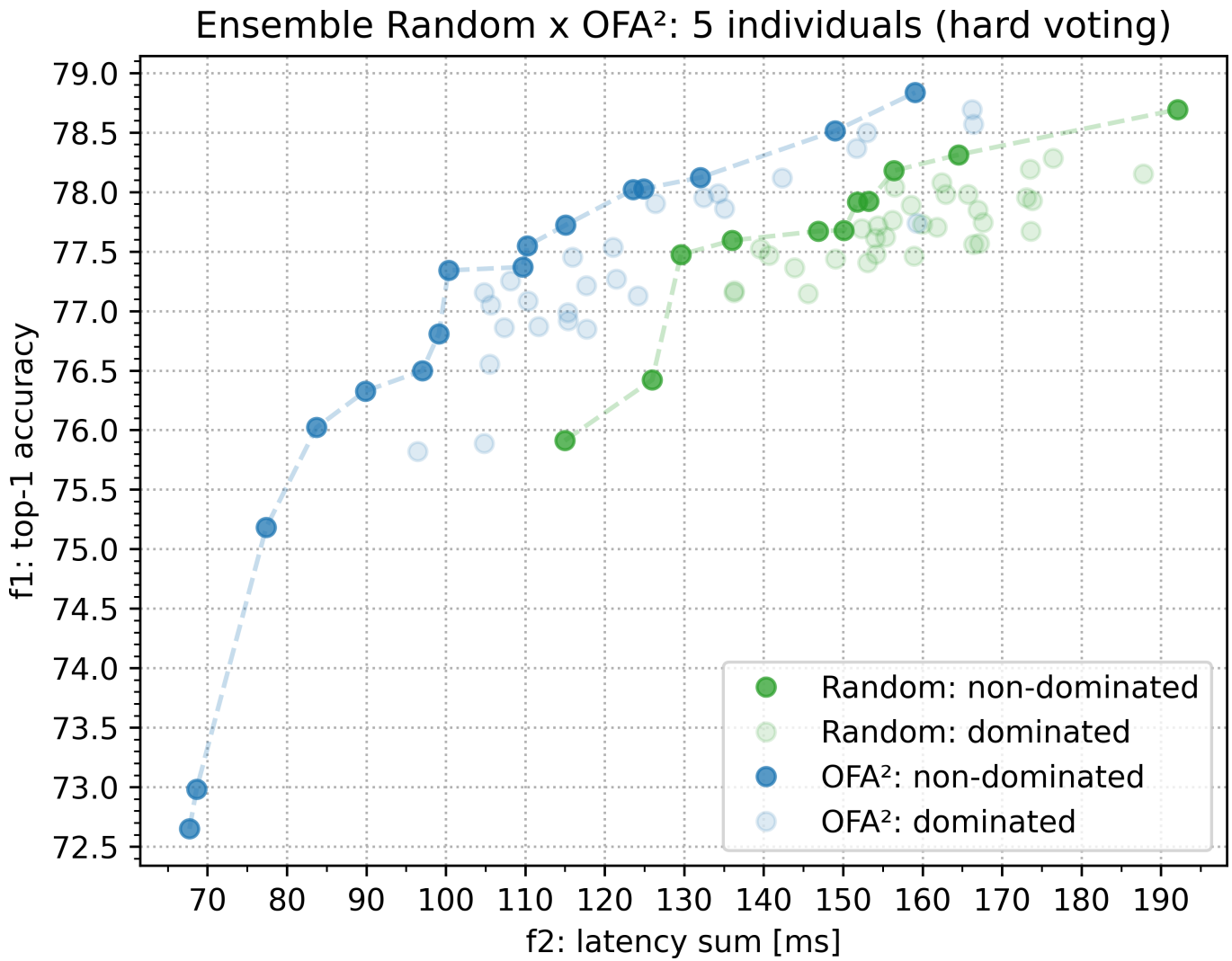
OFA² x Random

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - [4 individuals]
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



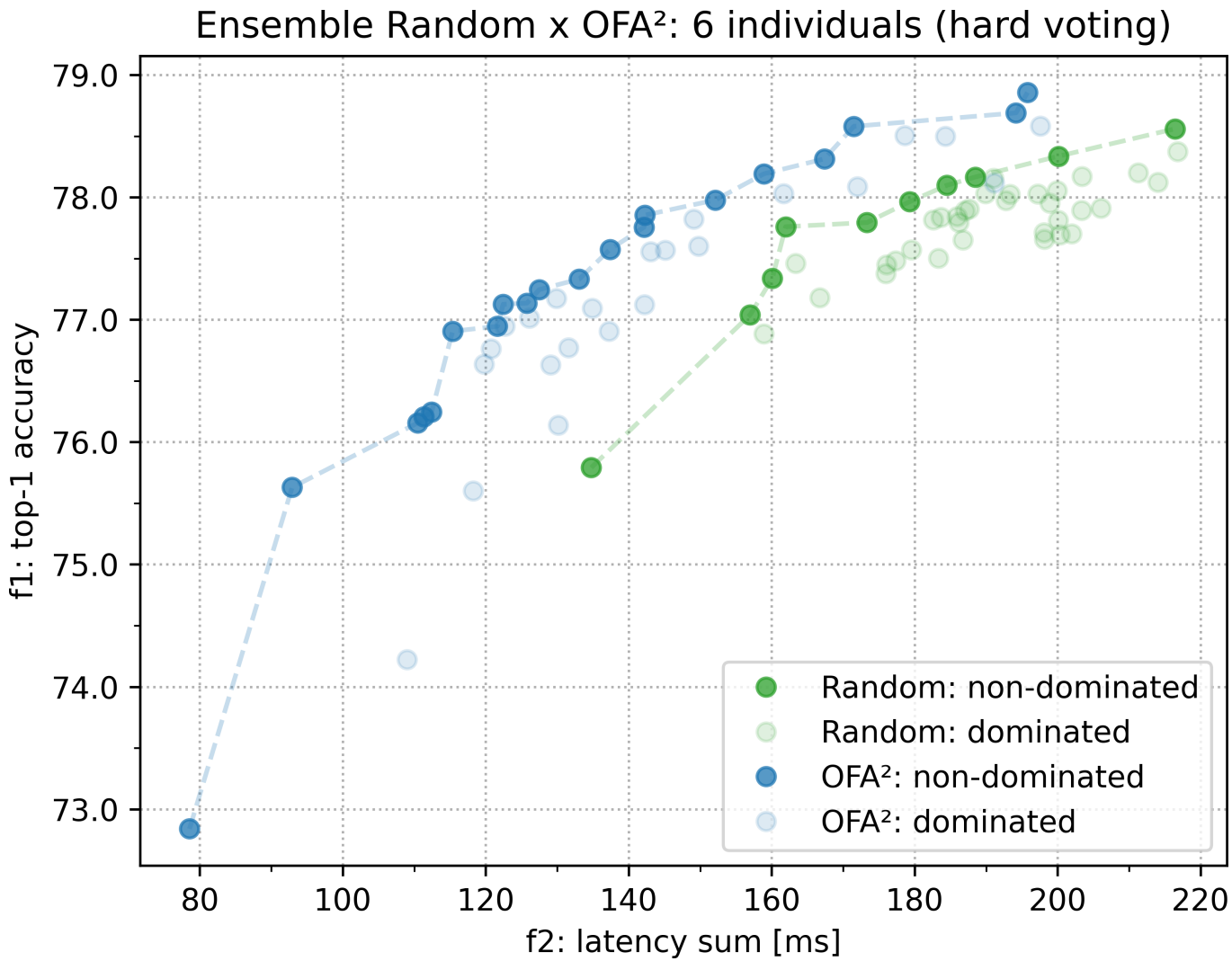
OFA² x Random

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - [5 individuals]
  - 6 individuals
  - 7 individuals
  - 8 individuals



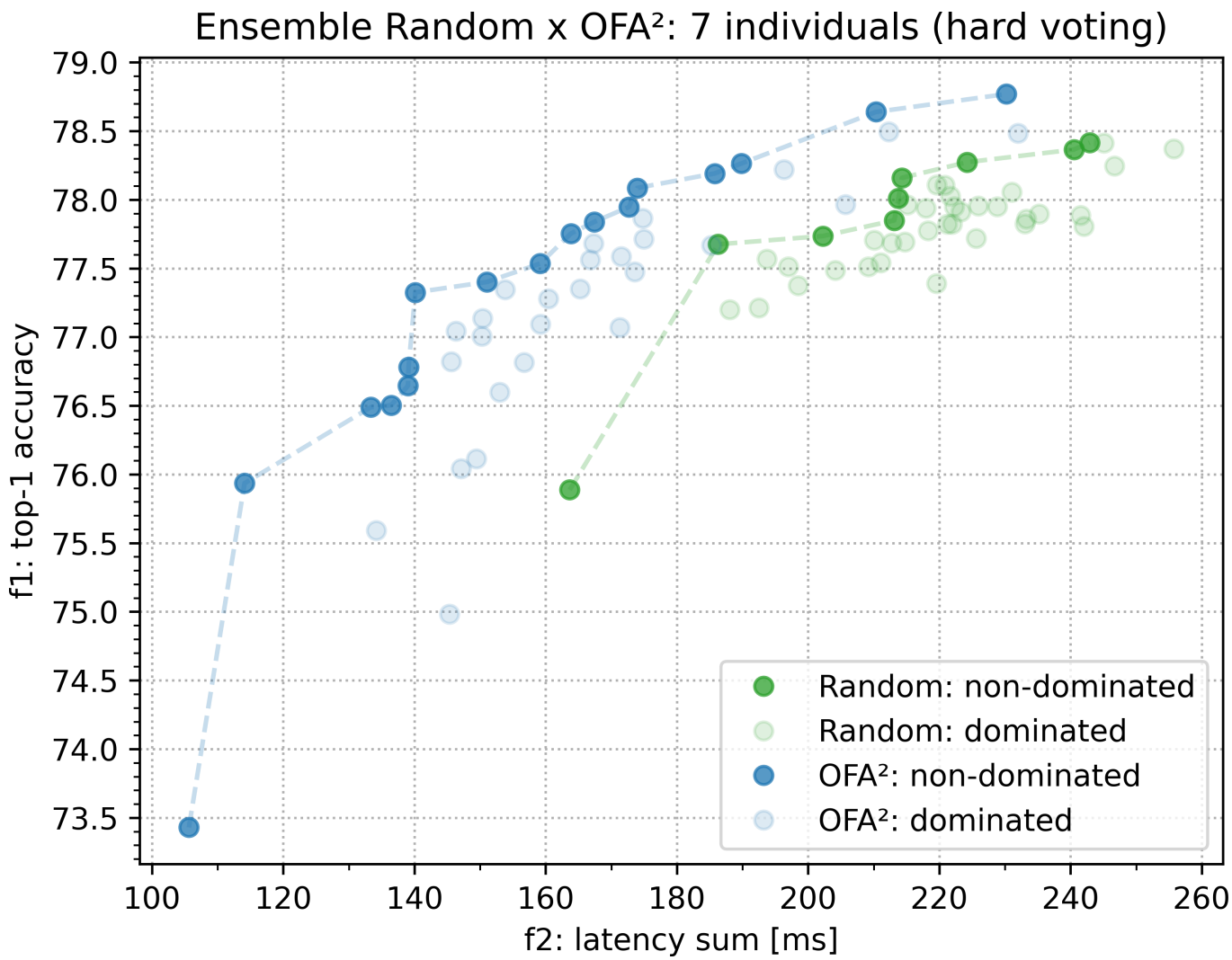
OFA² x Random

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - [6 individuals]
  - 7 individuals
  - 8 individuals



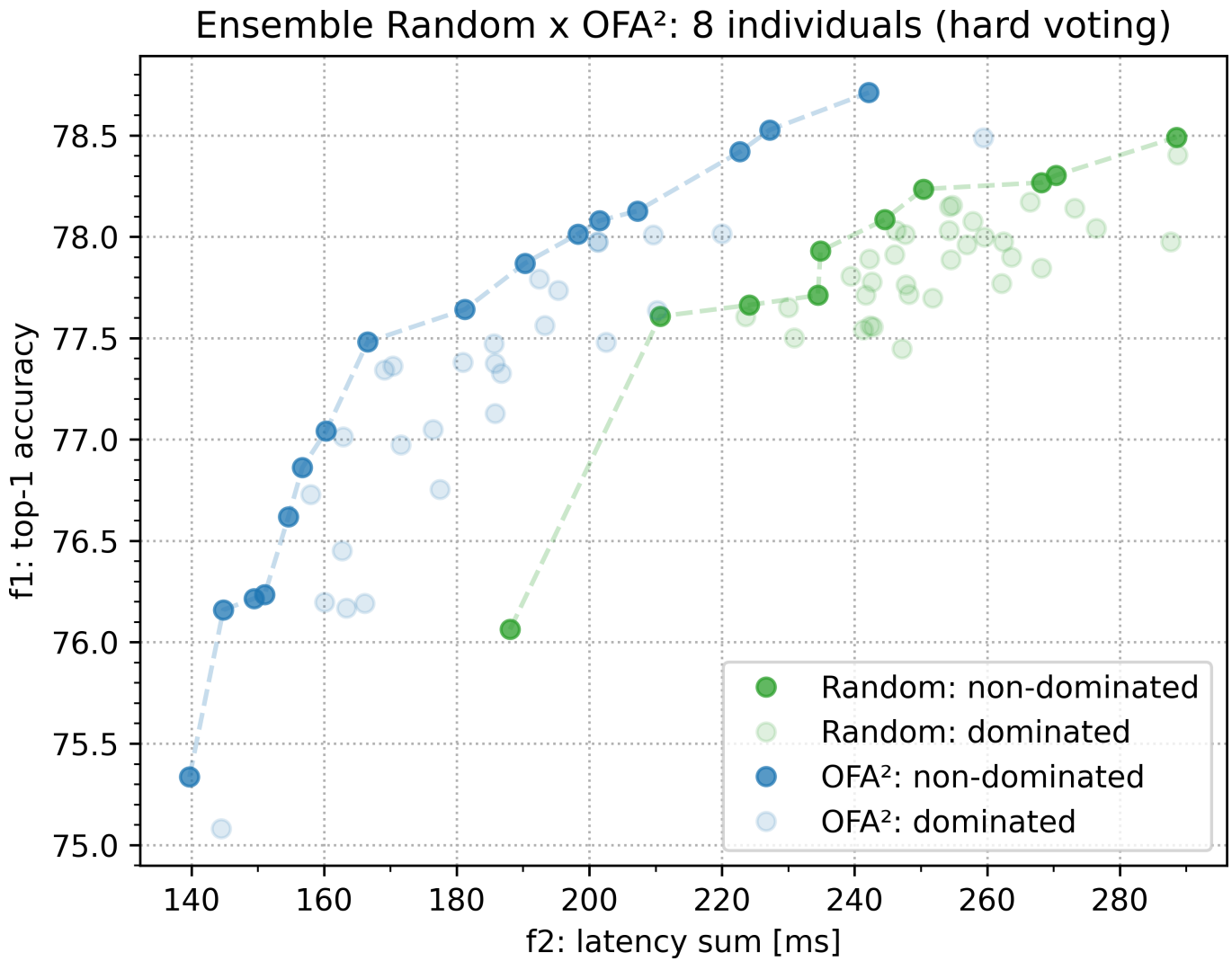
OFA² x Random

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - [7 individuals]
  - 8 individuals



OFA² x Random

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - [8 individuals]

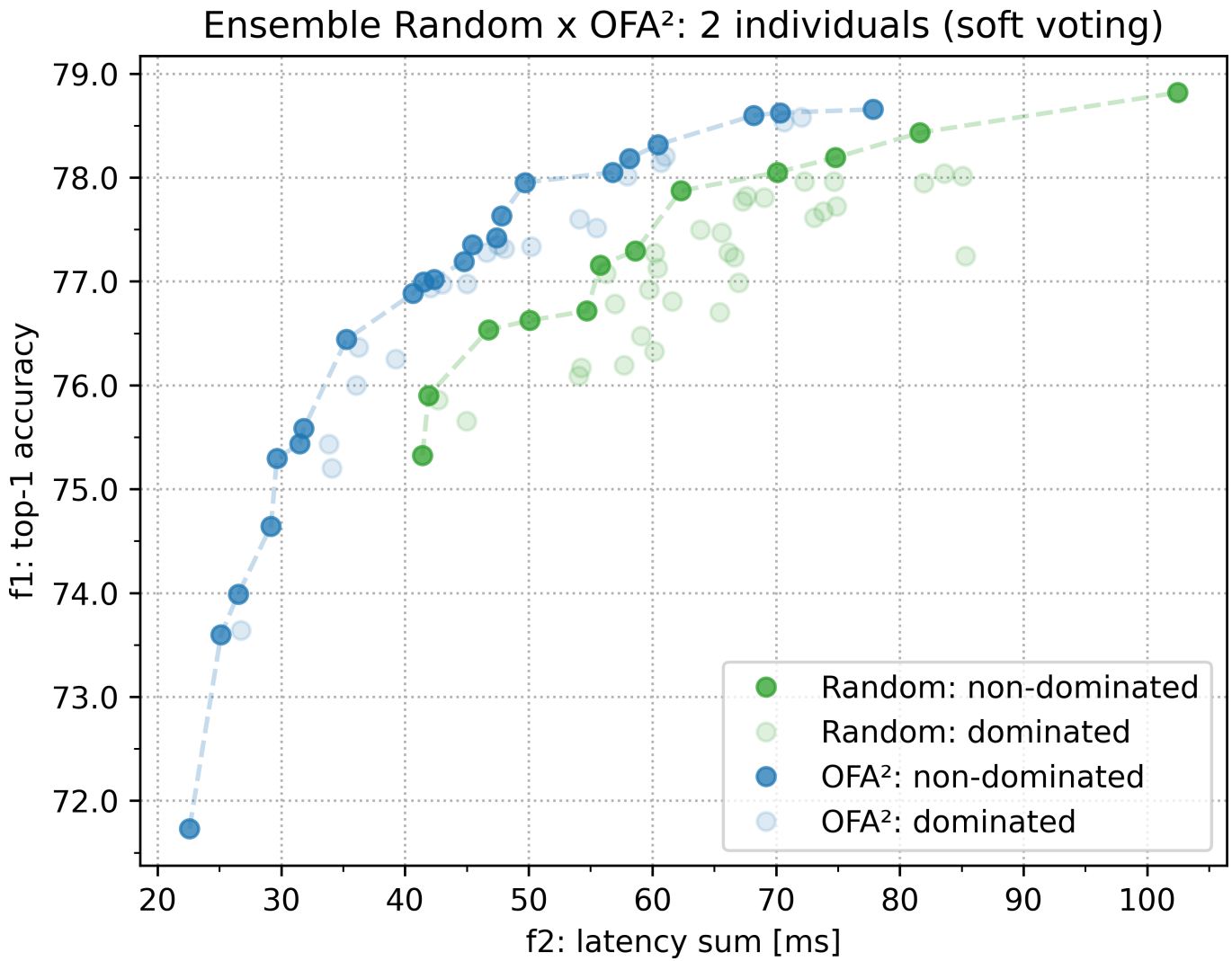


OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

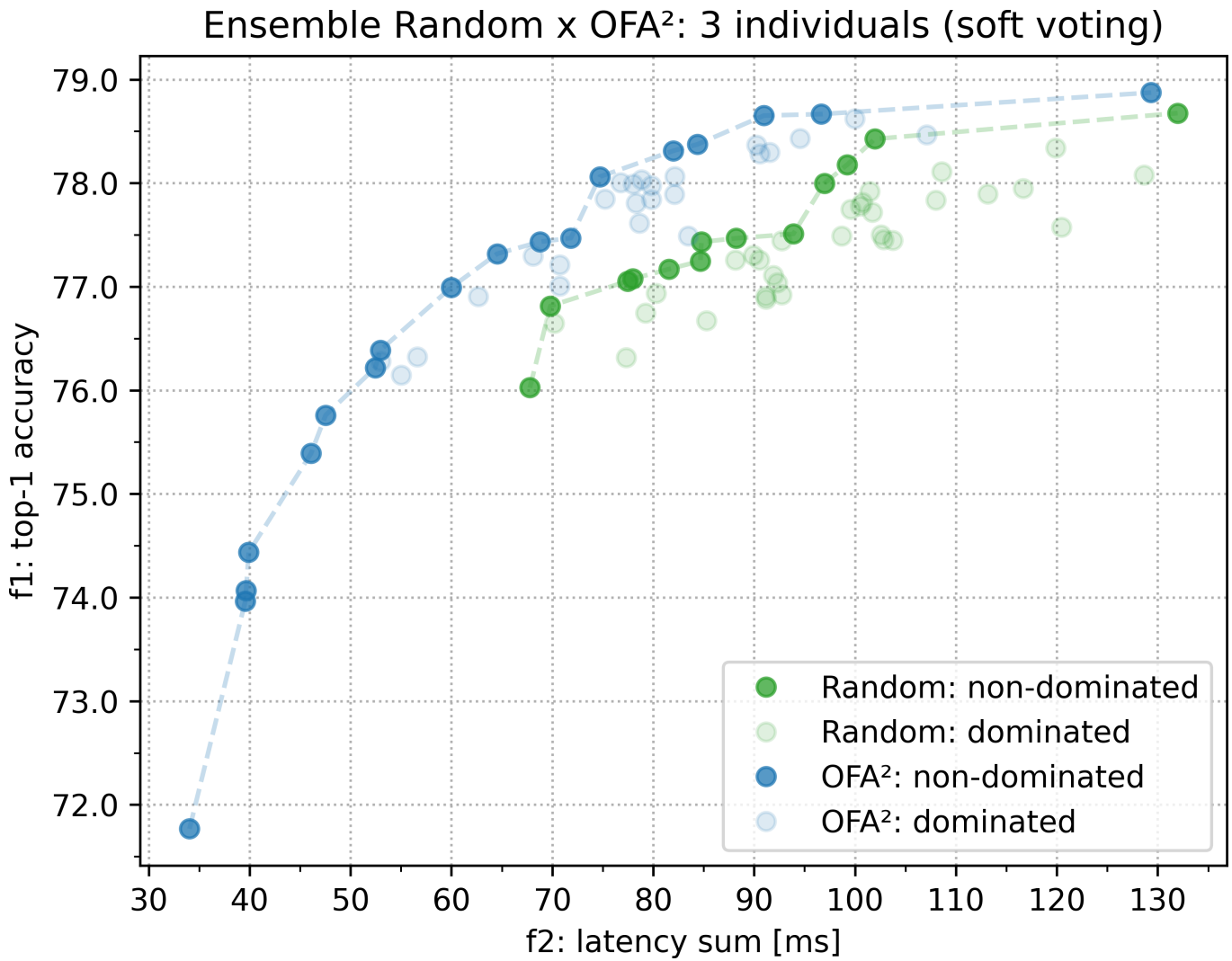
OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



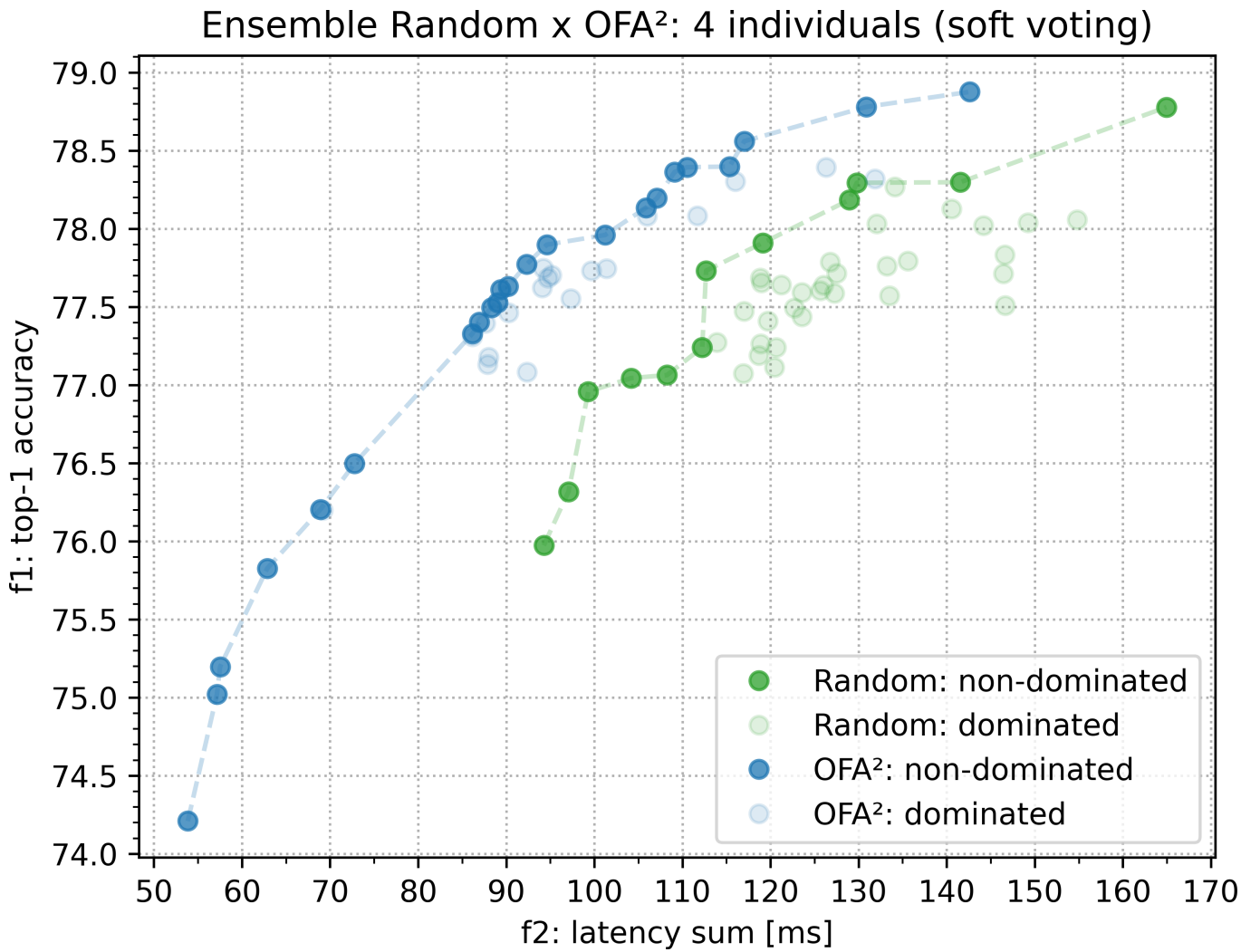
OFA² x Random

- **Voting**
  - hard
  - [soft]
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - [3 individuals]
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



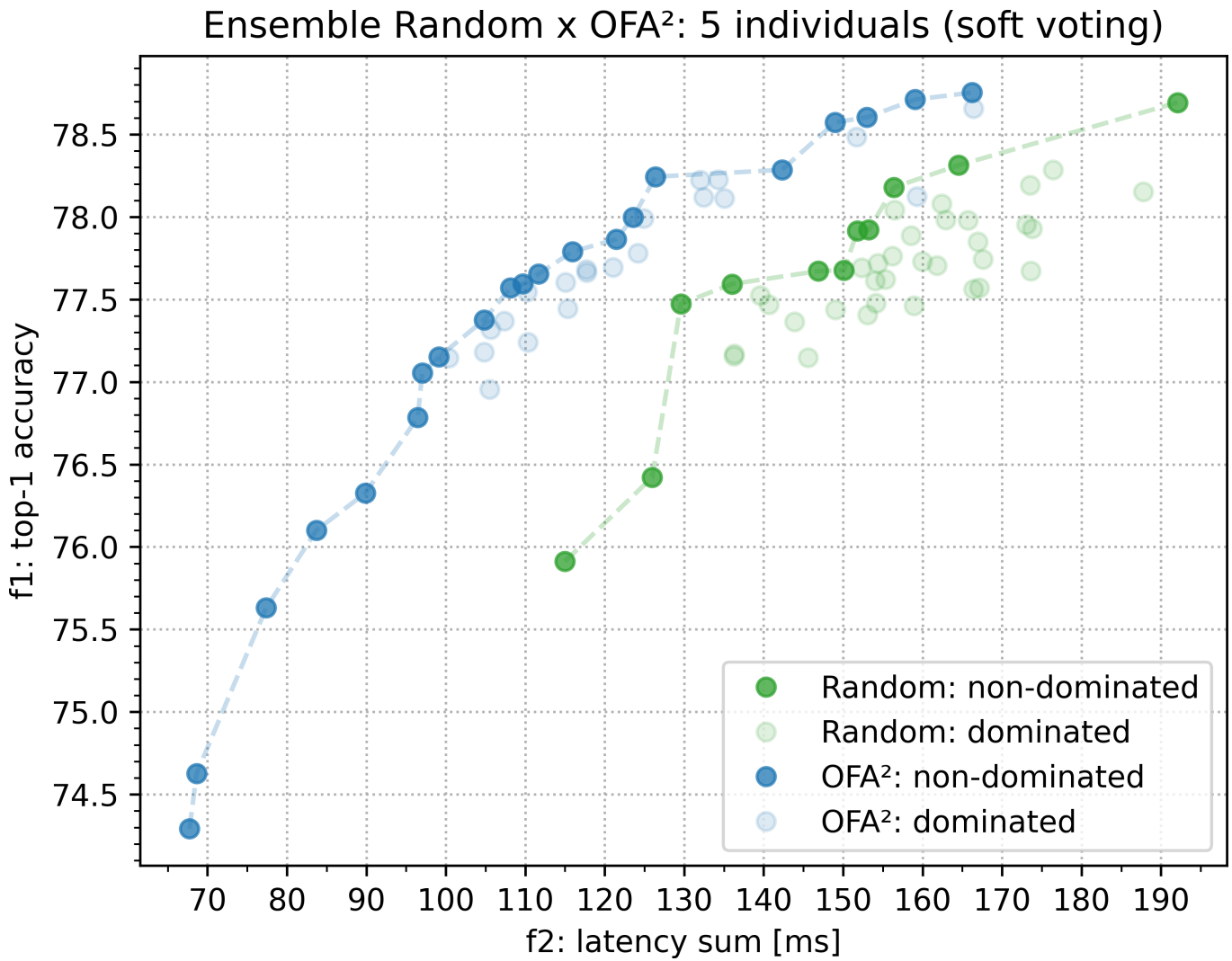
OFA² x Random

- **Voting**
  - hard
  - [soft]
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - [4 individuals]
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



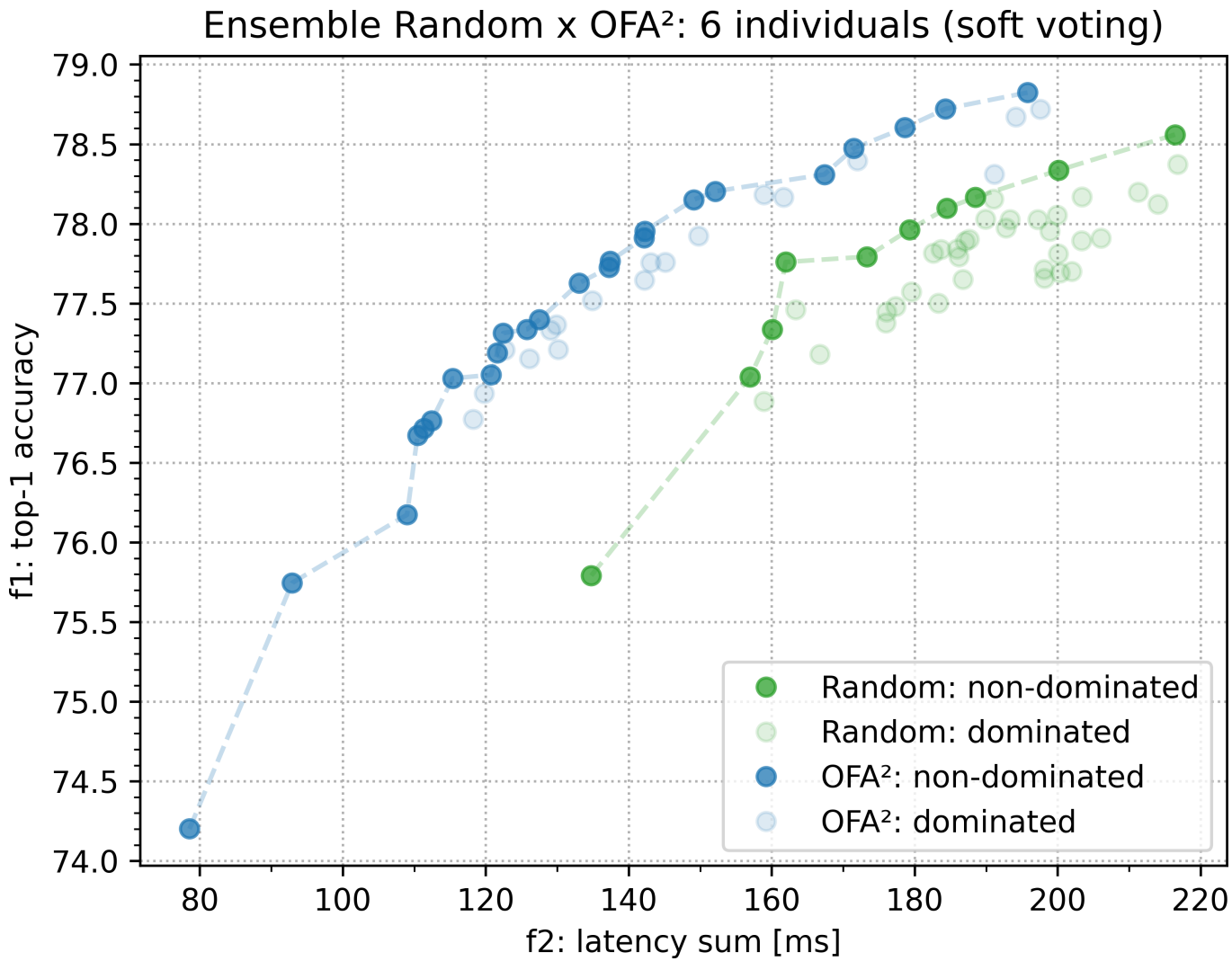
OFA² x Random

- **Voting**
  - hard
  - [soft]
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - [5 individuals]
  - 6 individuals
  - 7 individuals
  - 8 individuals



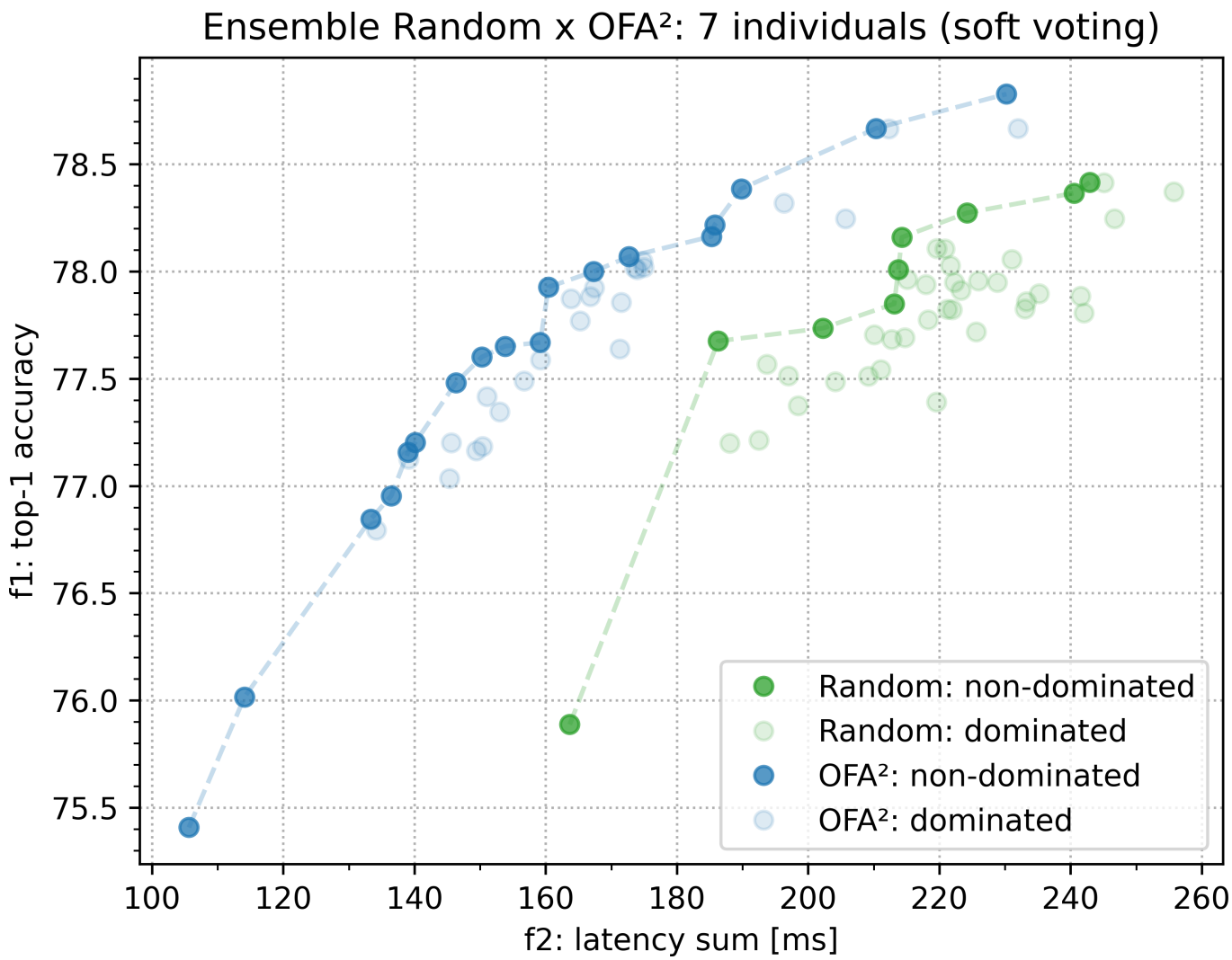
OFA² x Random

- **Voting**
  - hard
  - [soft]
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - [6 individuals]
  - 7 individuals
  - 8 individuals



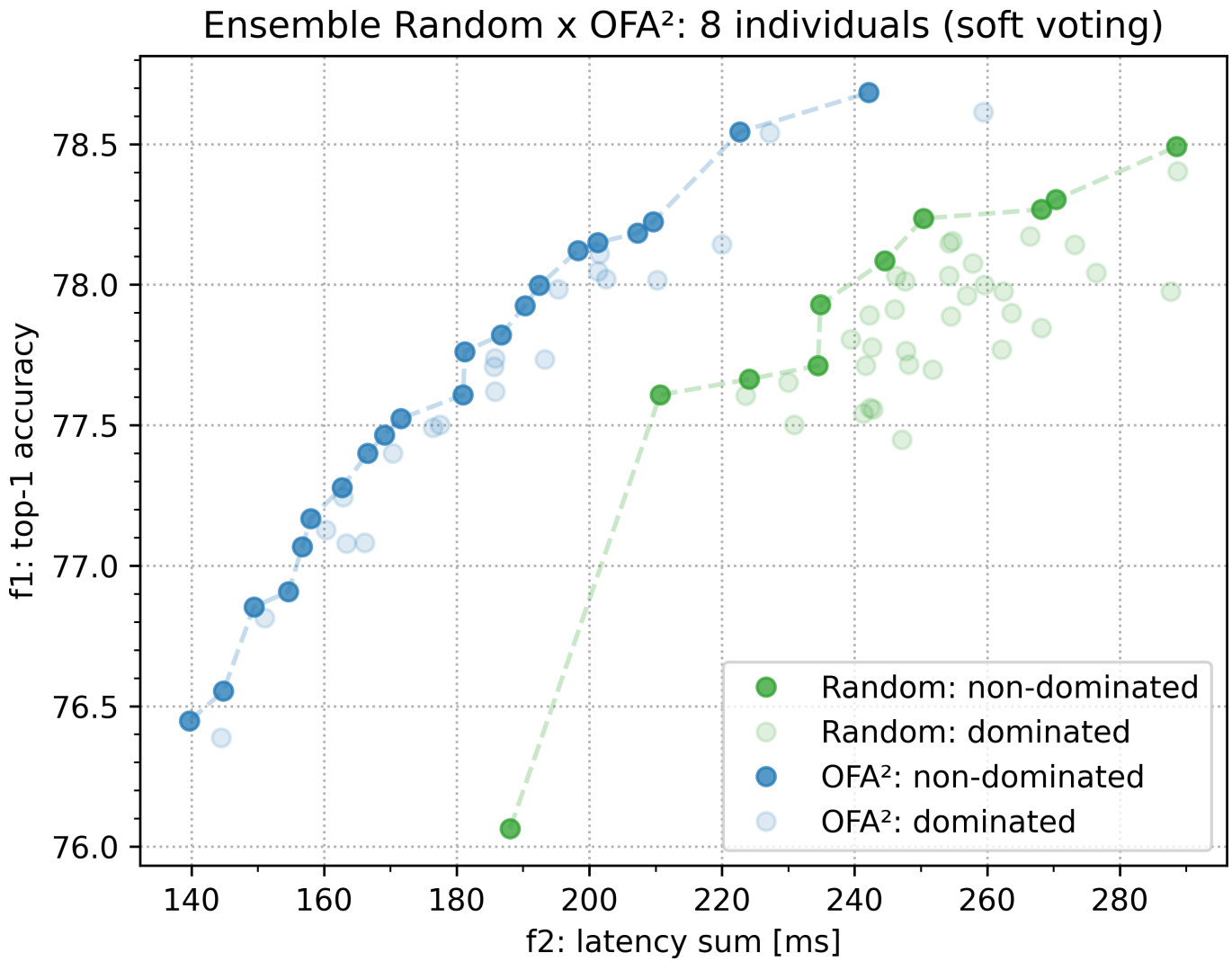
OFA² x Random

- **Voting**
  - hard
  - [soft]
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - [7 individuals]
  - 8 individuals



OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

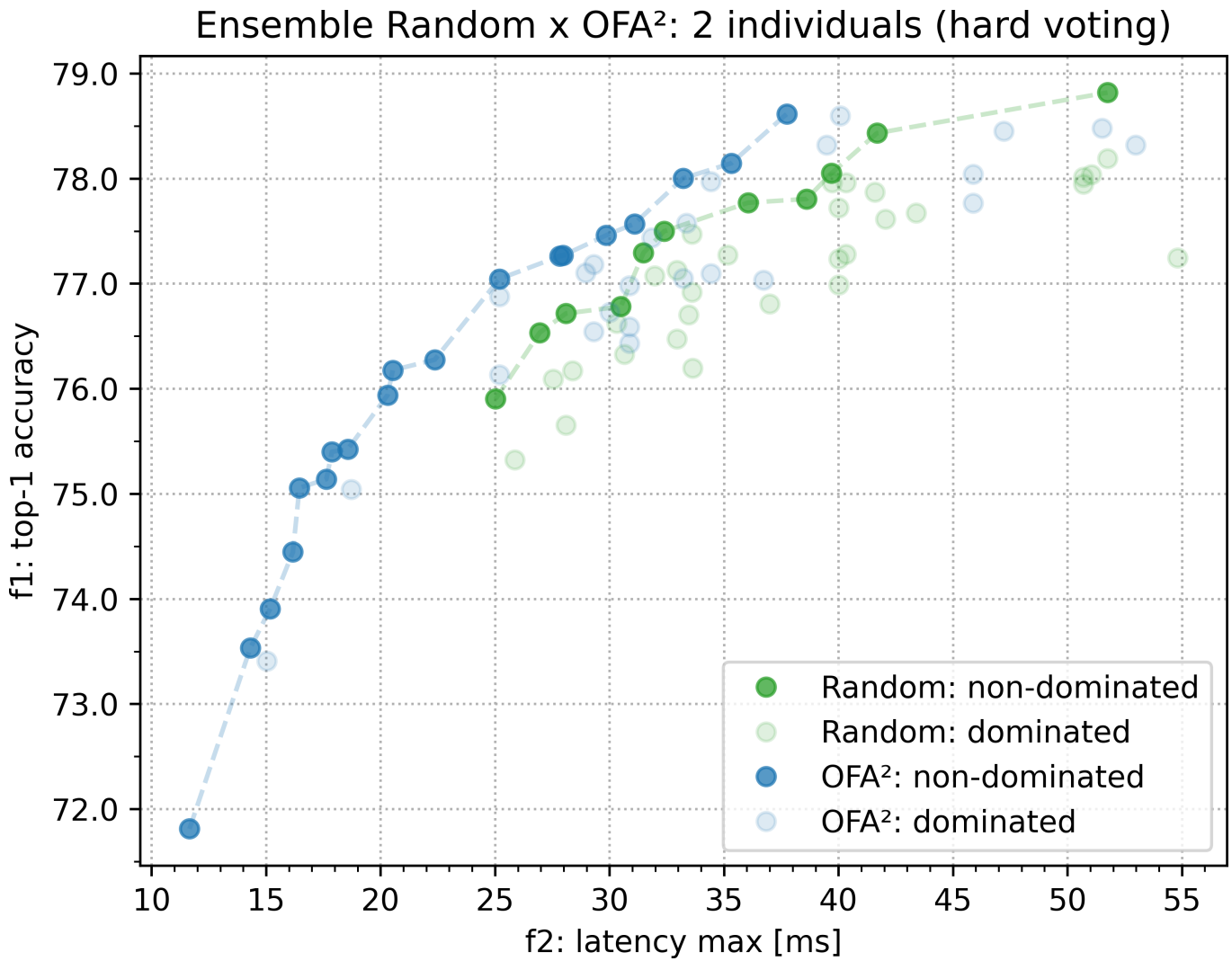


OFA² x Random

- **Voting**
  - [hard]
  - soft
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

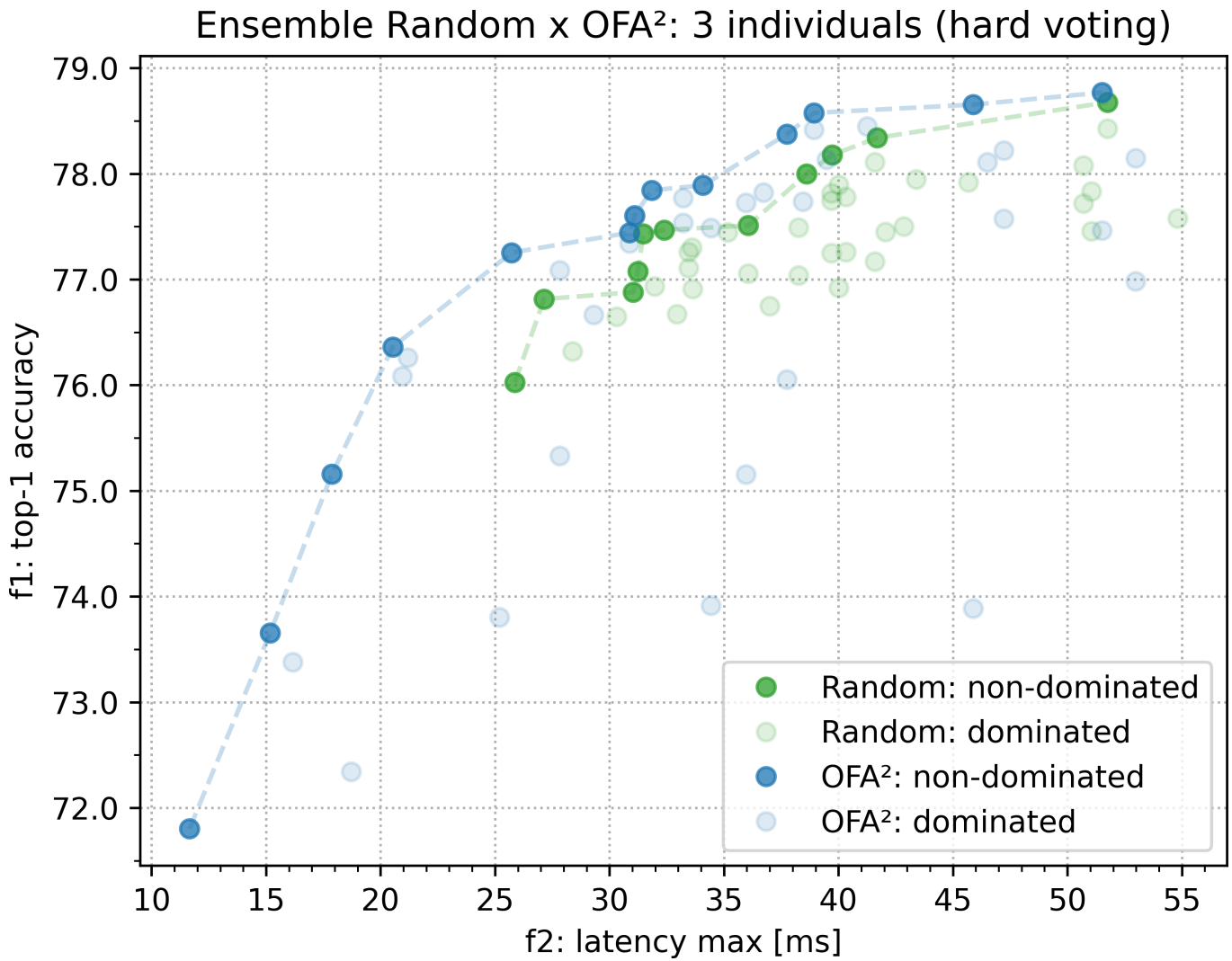
OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



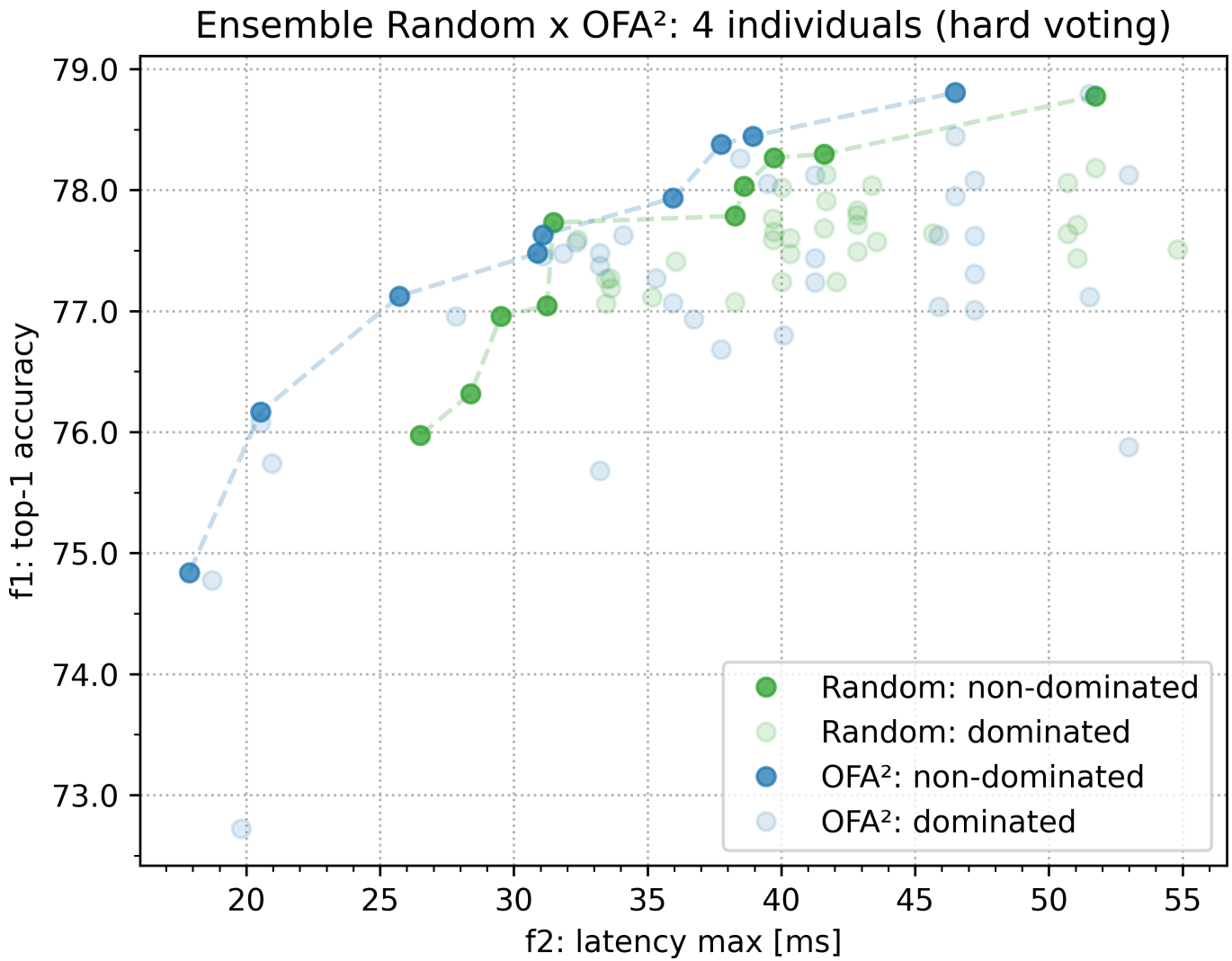
OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



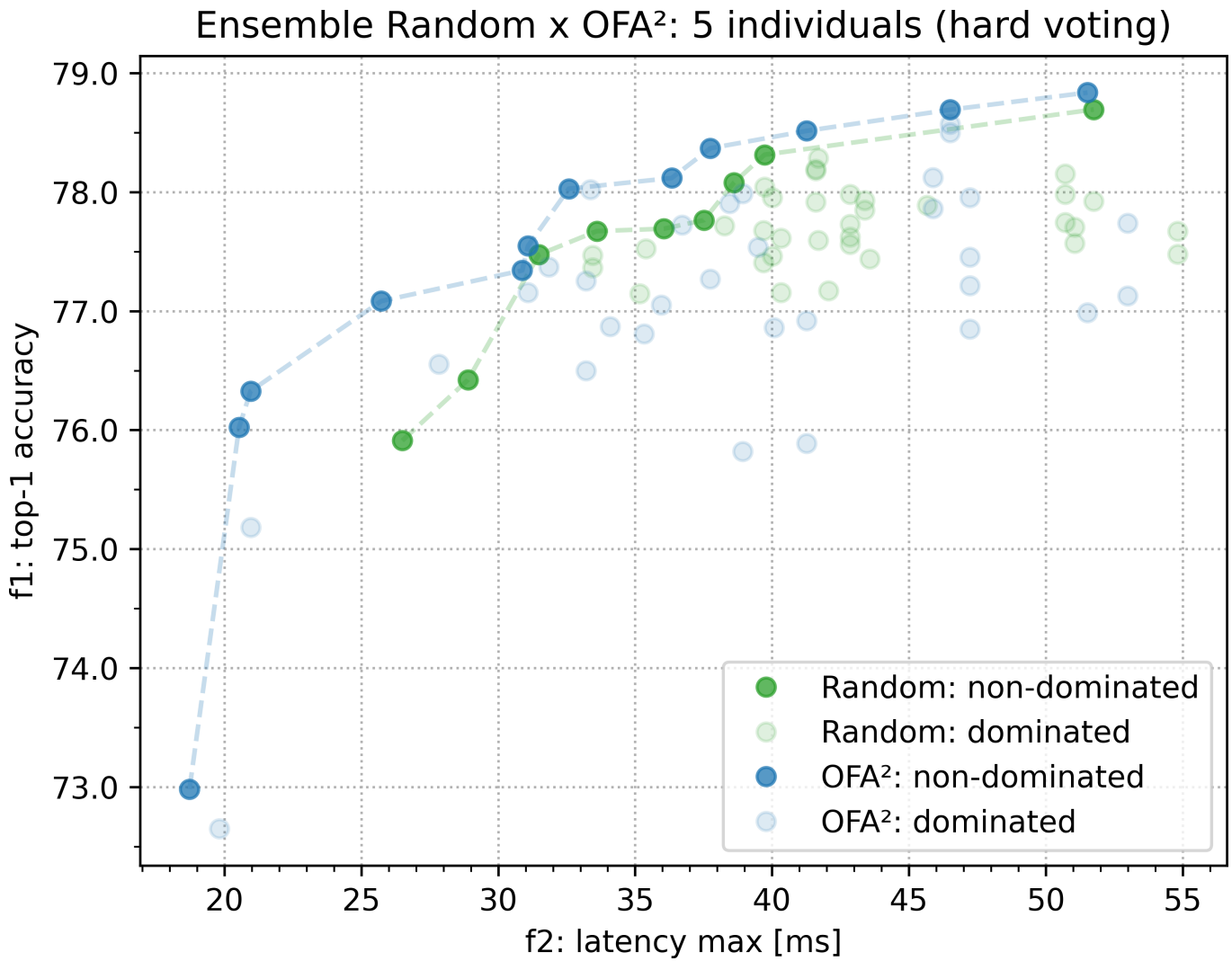
OFA² x Random

- **Voting**
  - [hard]
  - soft
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - [4 individuals]
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



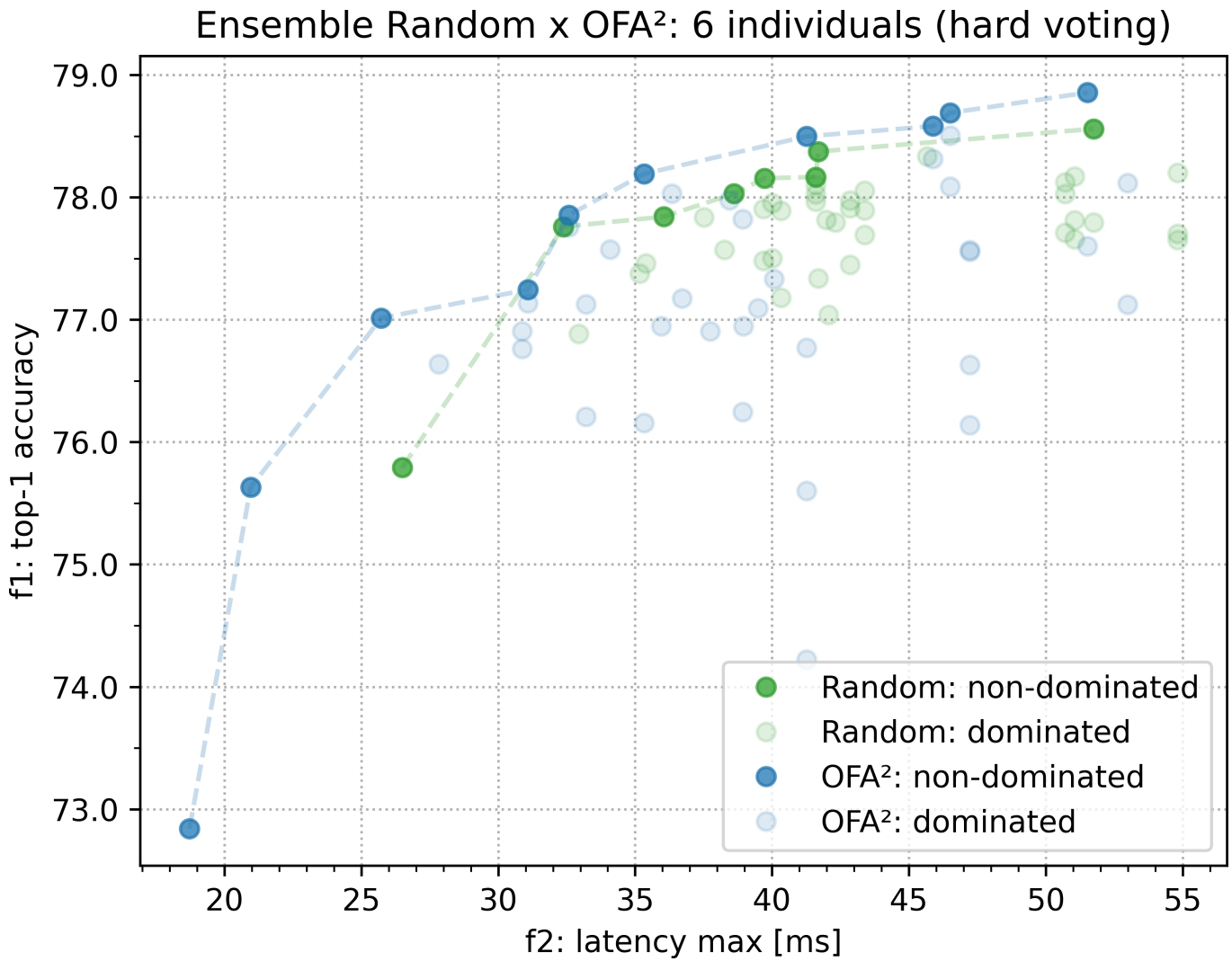
OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



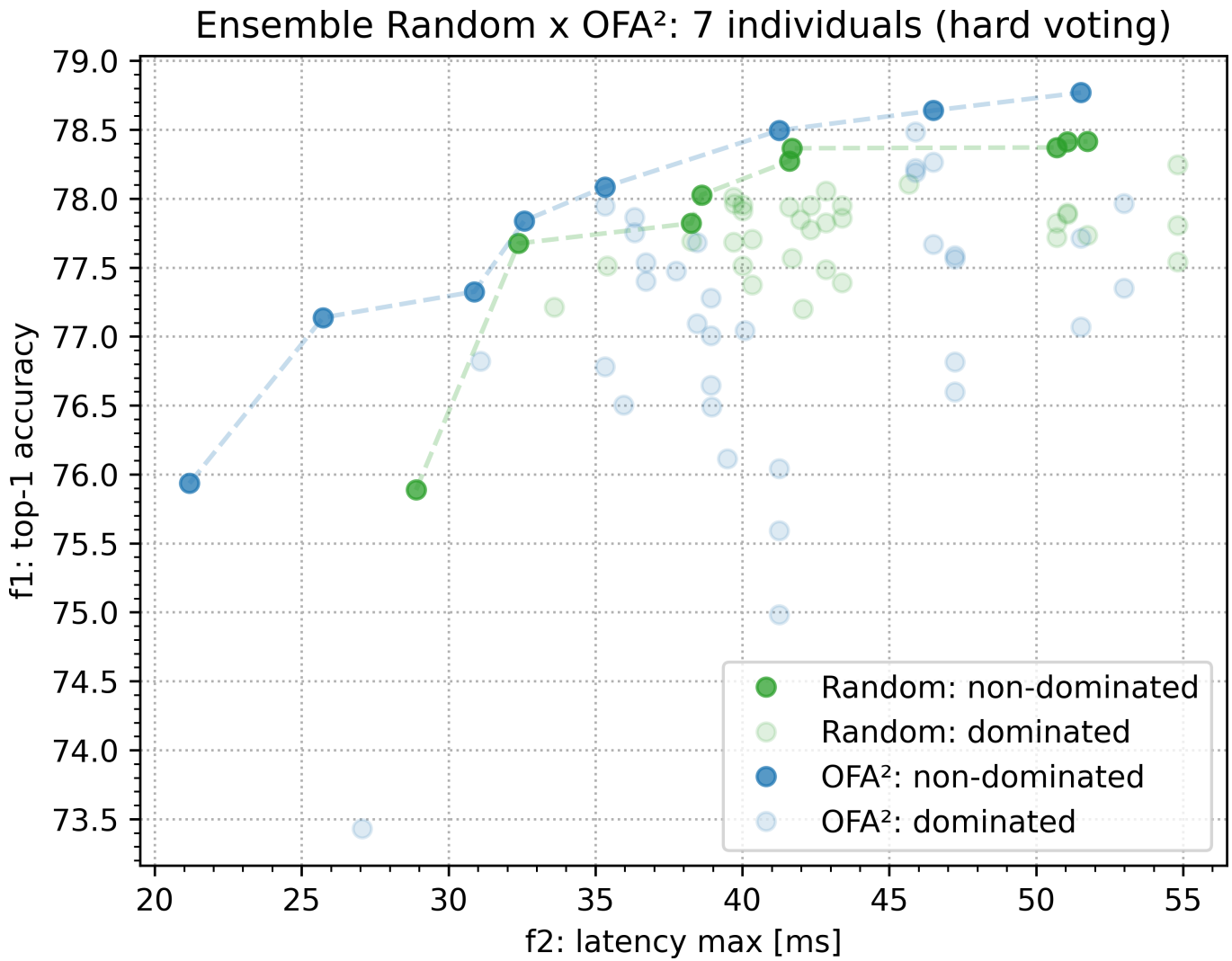
OFA² x Random

- **Voting**
  - [hard]
  - soft
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - [6 individuals]
  - 7 individuals
  - 8 individuals



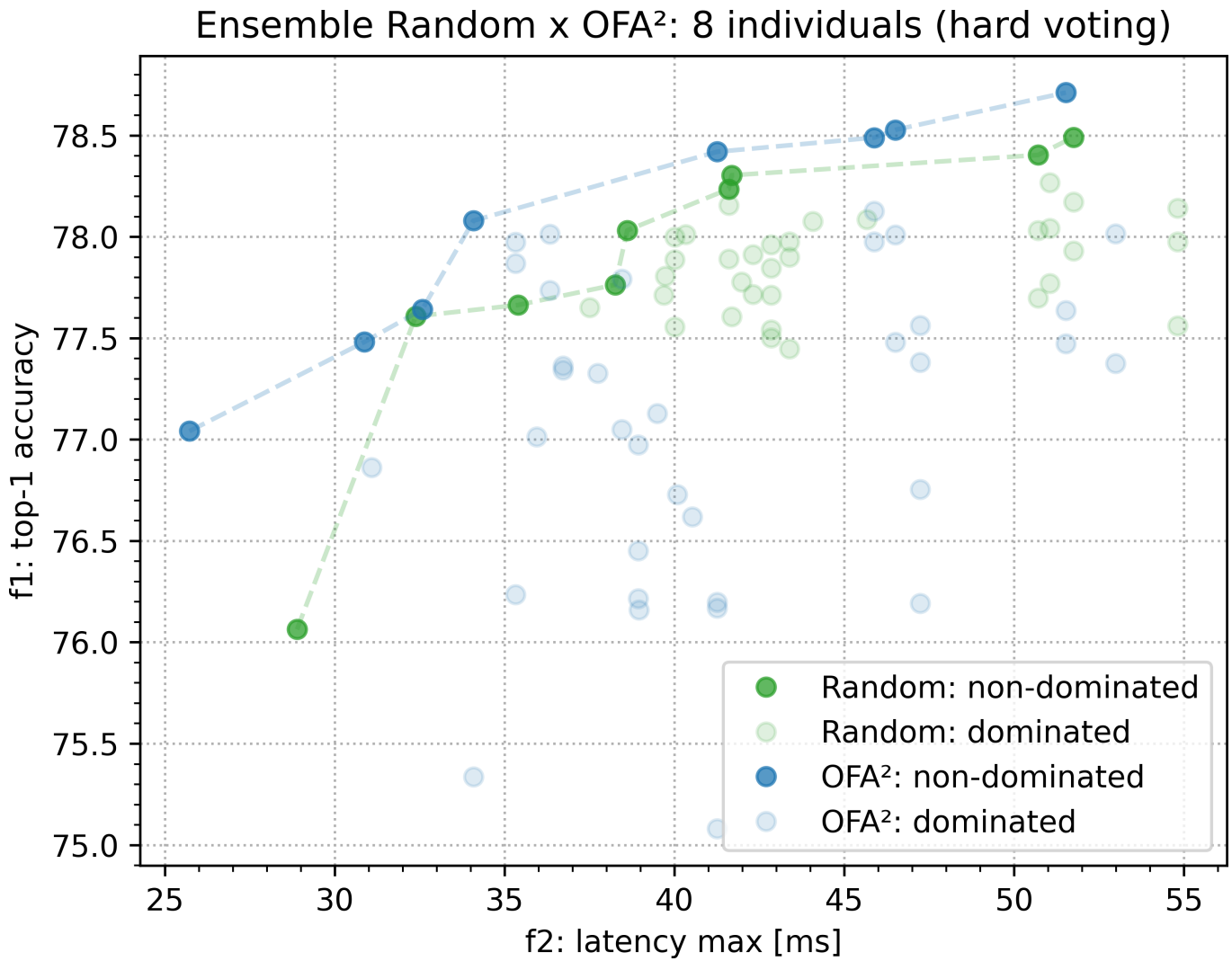
OFA² x Random

- **Voting**
  - [hard]
  - soft
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - [7 individuals]
  - 8 individuals



OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

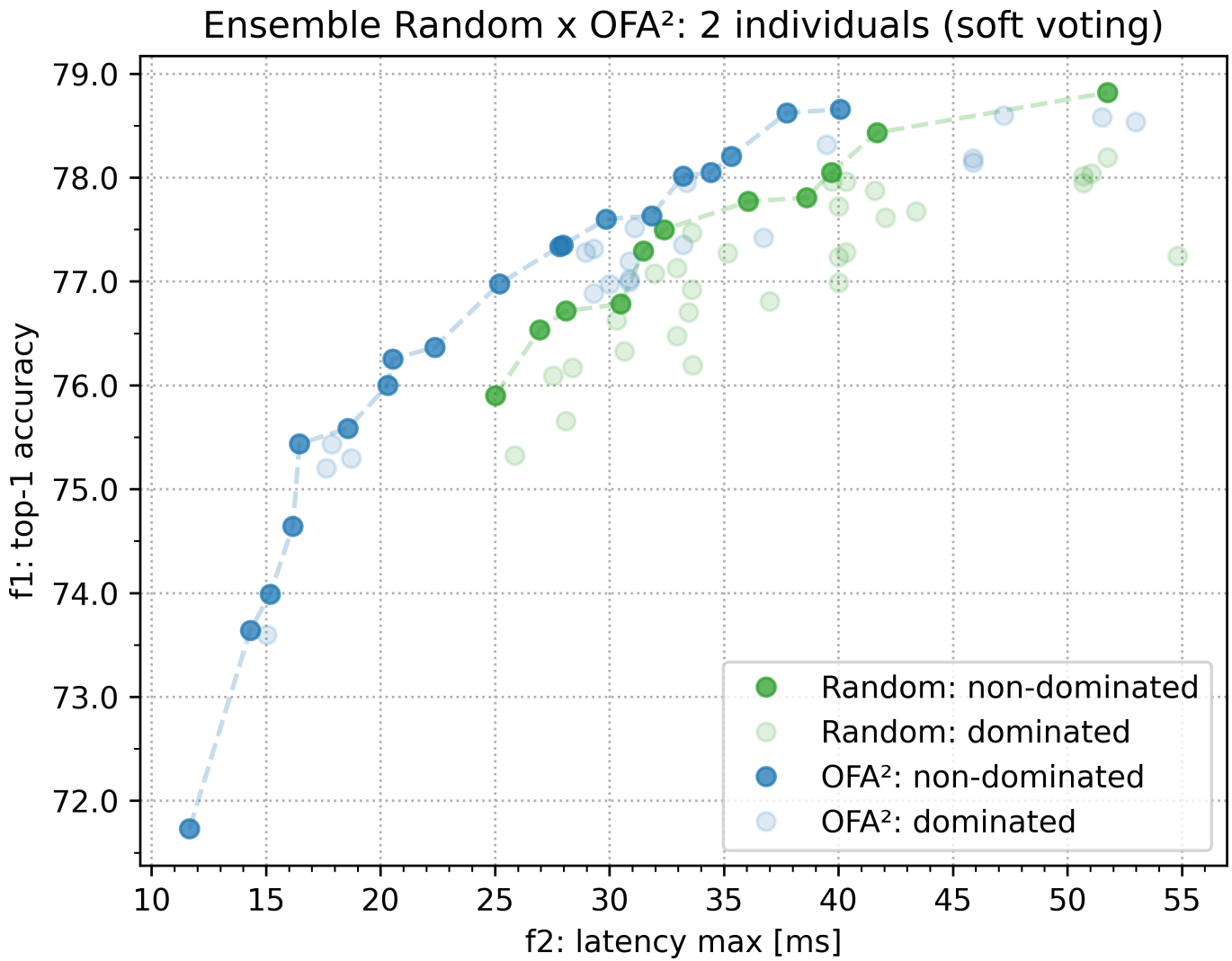


OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

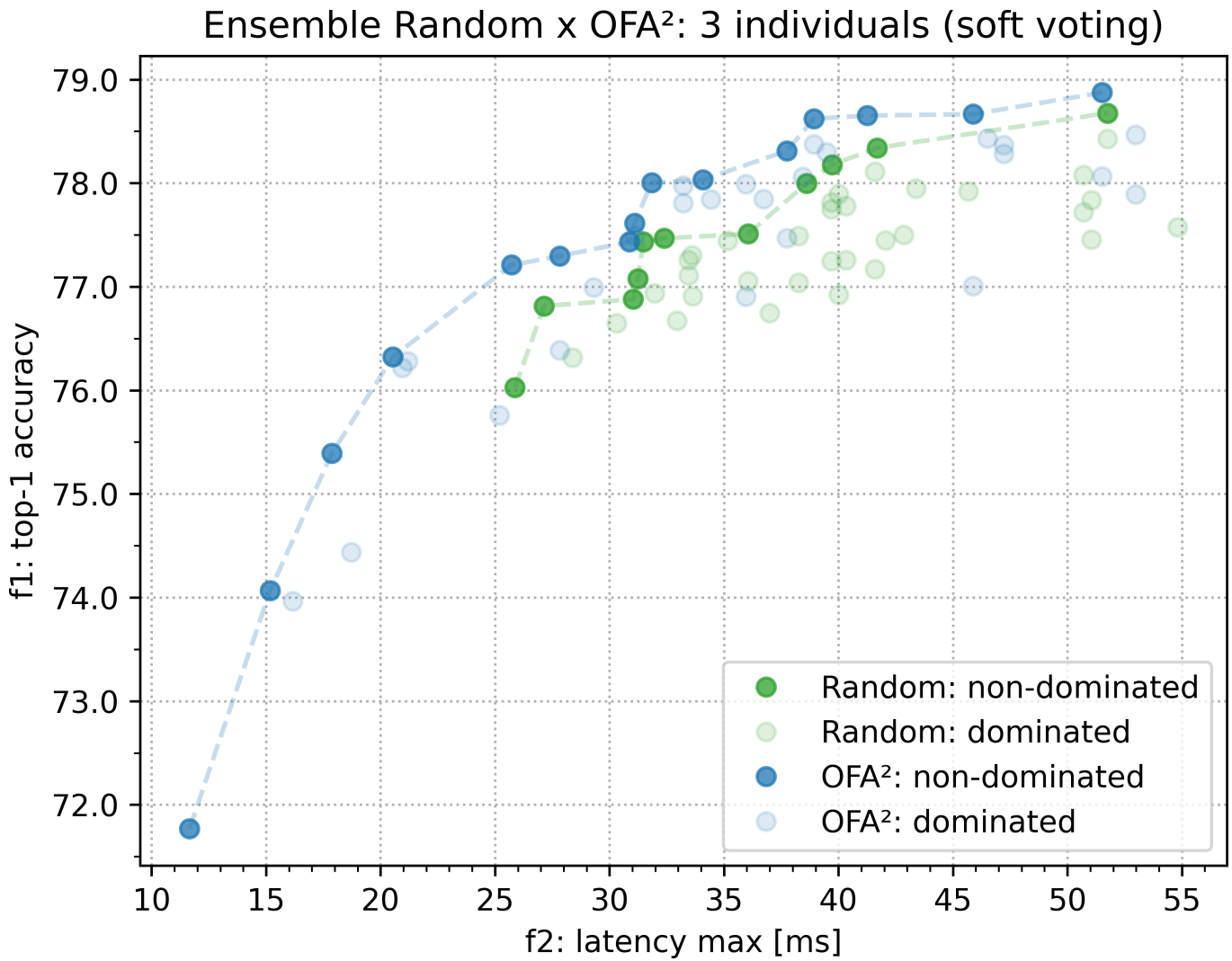
OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



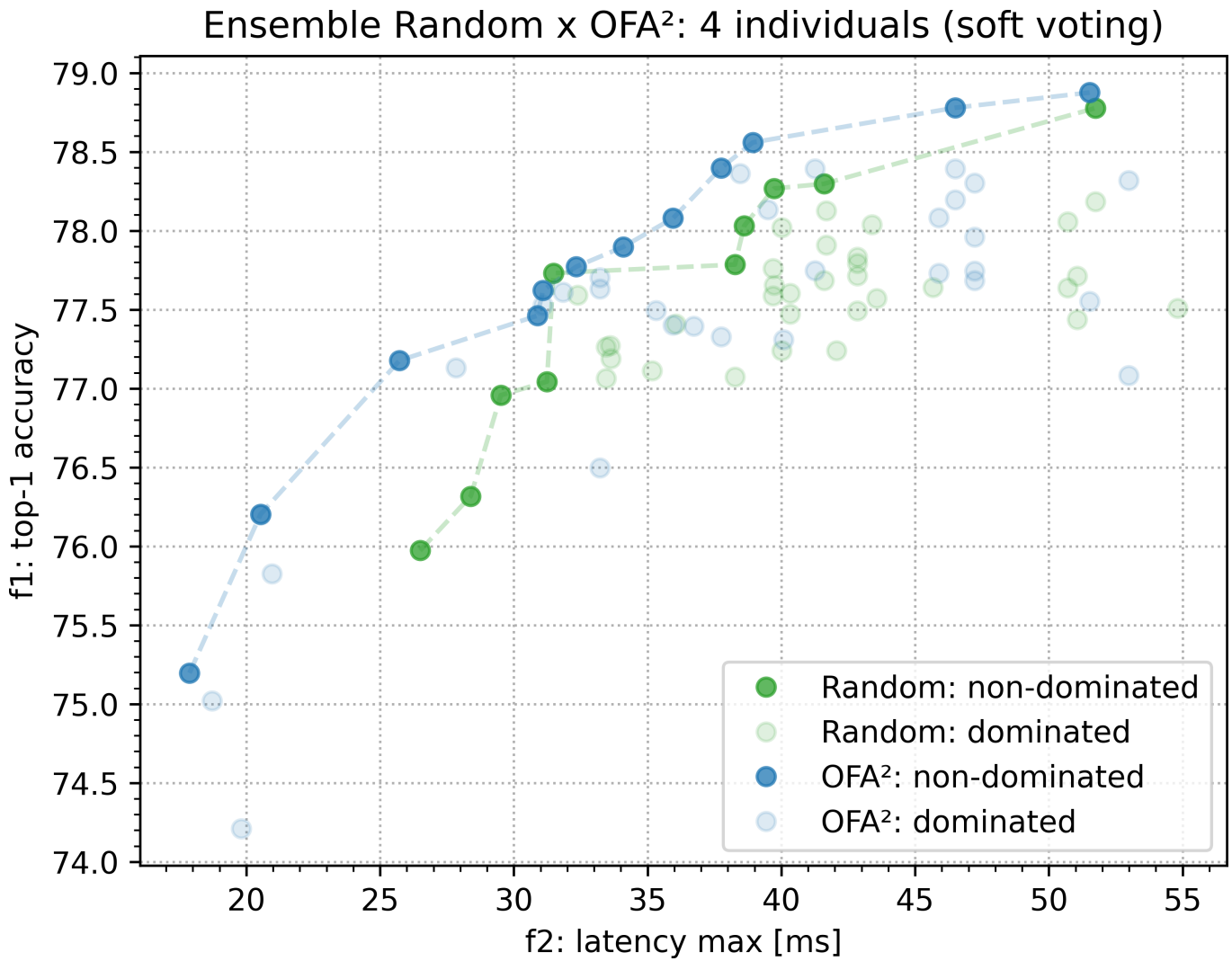
OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



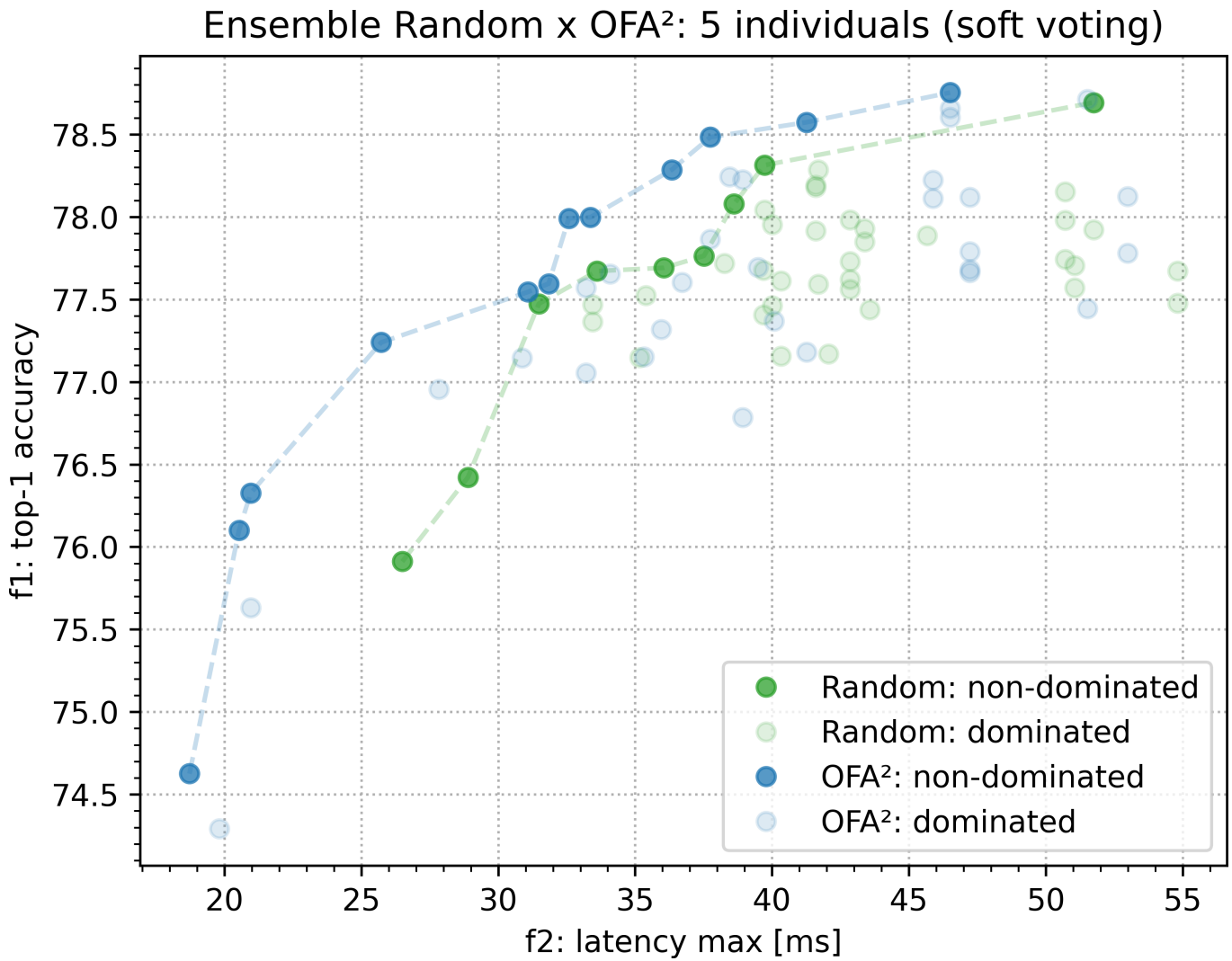
OFA² x Random

- **Voting**
  - hard
  - [soft]
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - [4 individuals]
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



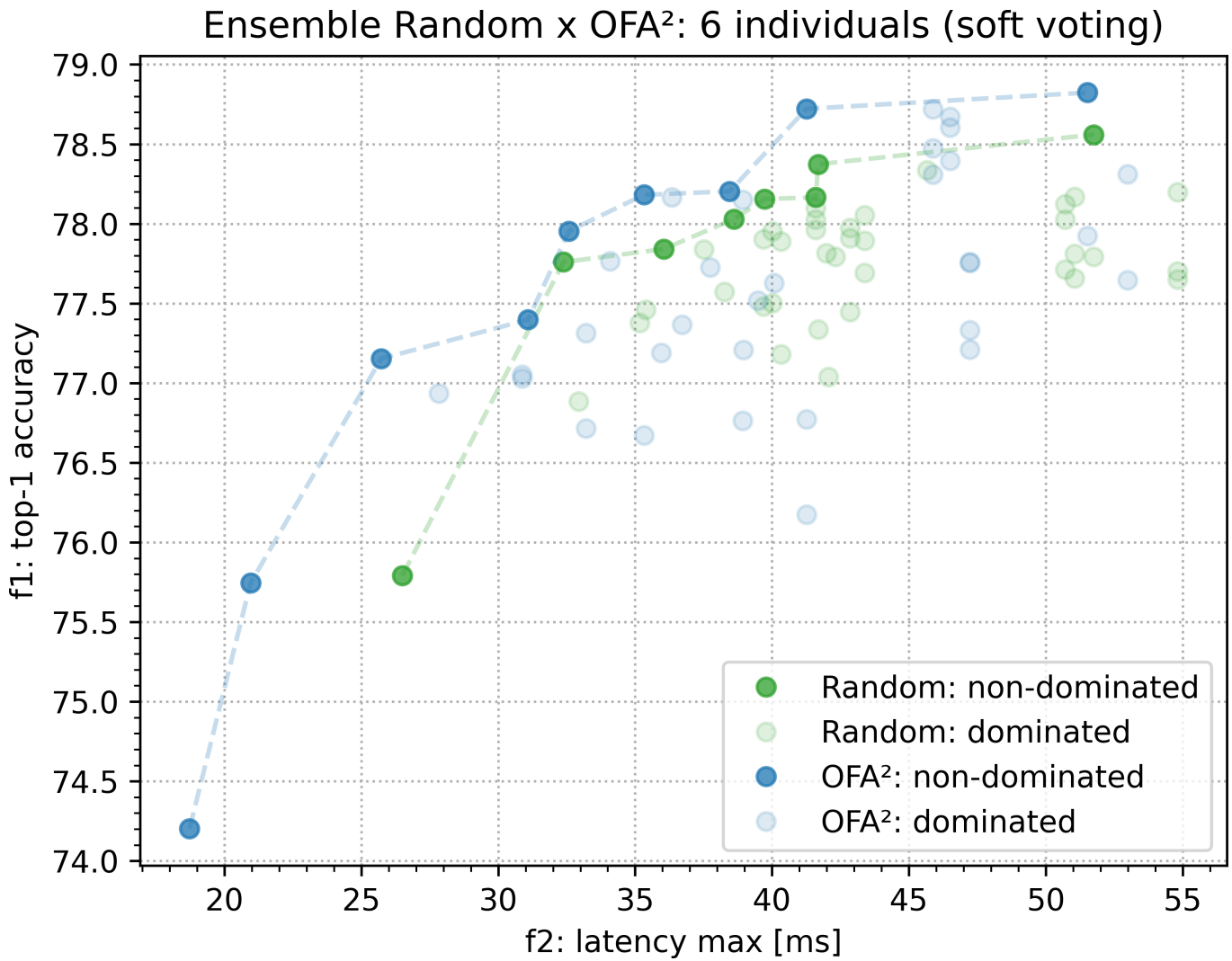
OFA² x Random

- **Voting**
  - hard
  - [soft]
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - [5 individuals]
  - 6 individuals
  - 7 individuals
  - 8 individuals



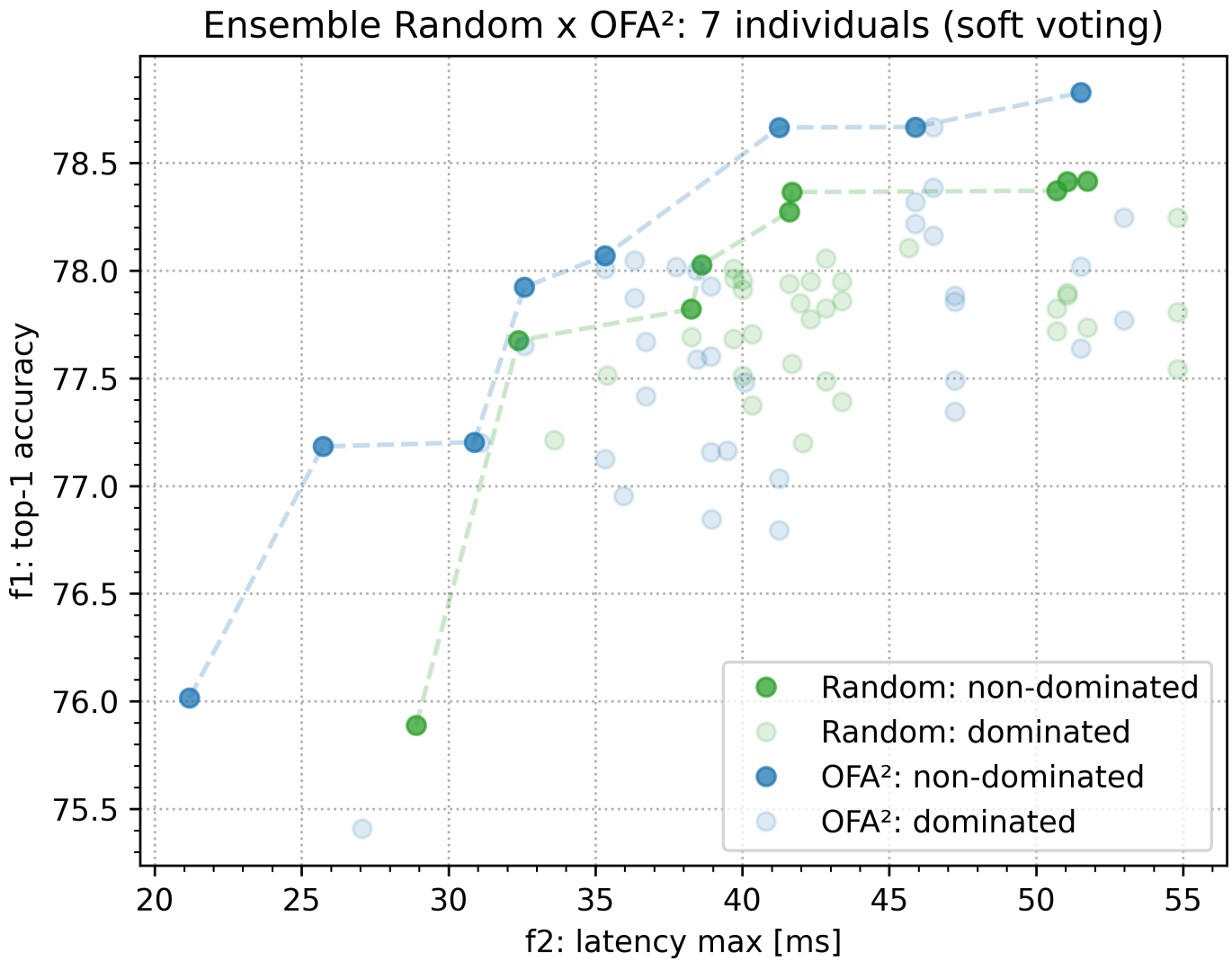
OFA² x Random

- **Voting**
  - hard
  - [soft]
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - [6 individuals]
  - 7 individuals
  - 8 individuals



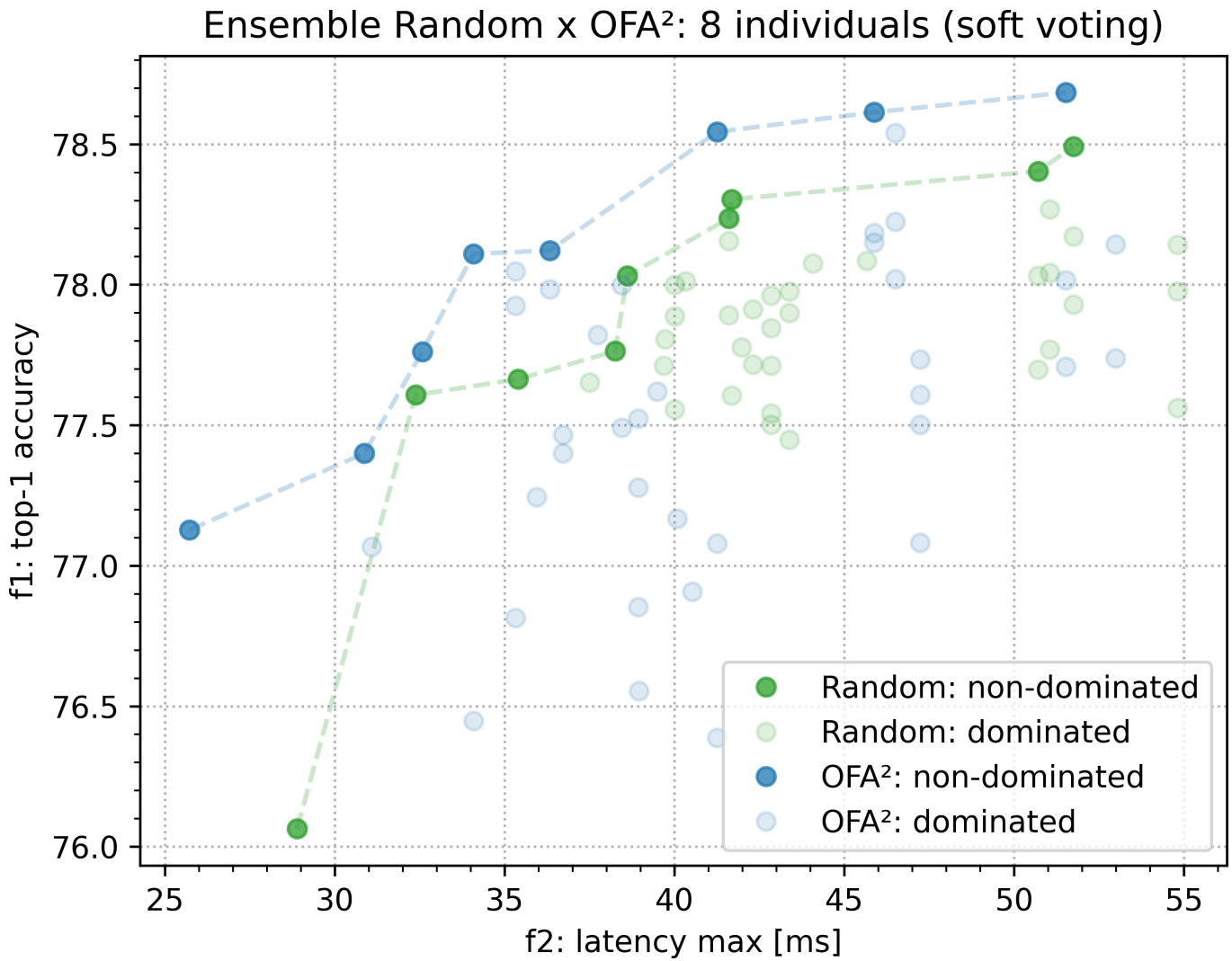
OFA² x Random

- **Voting**
  - hard
  - [soft]
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - [7 individuals]
  - 8 individuals



OFA² x Random

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

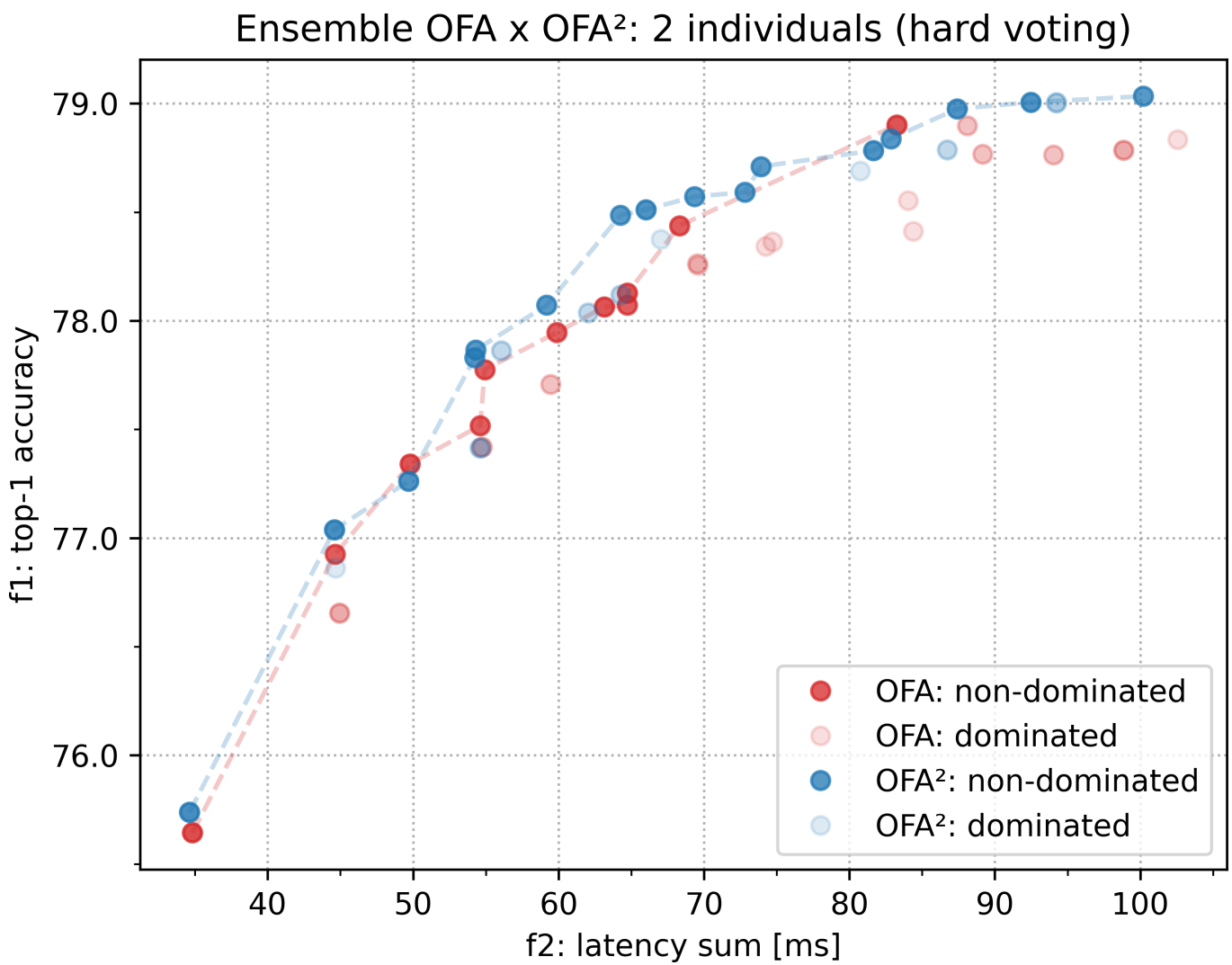


OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

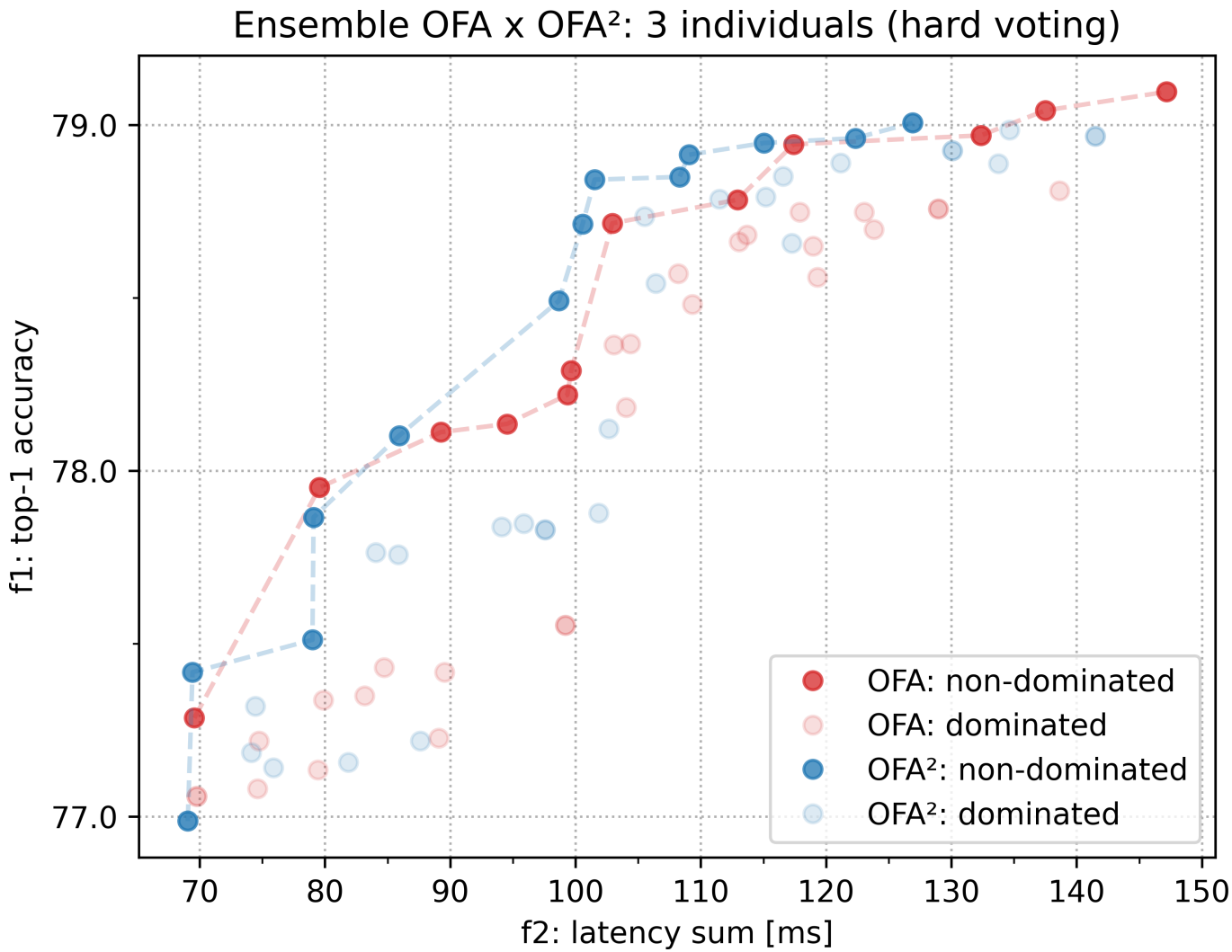
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



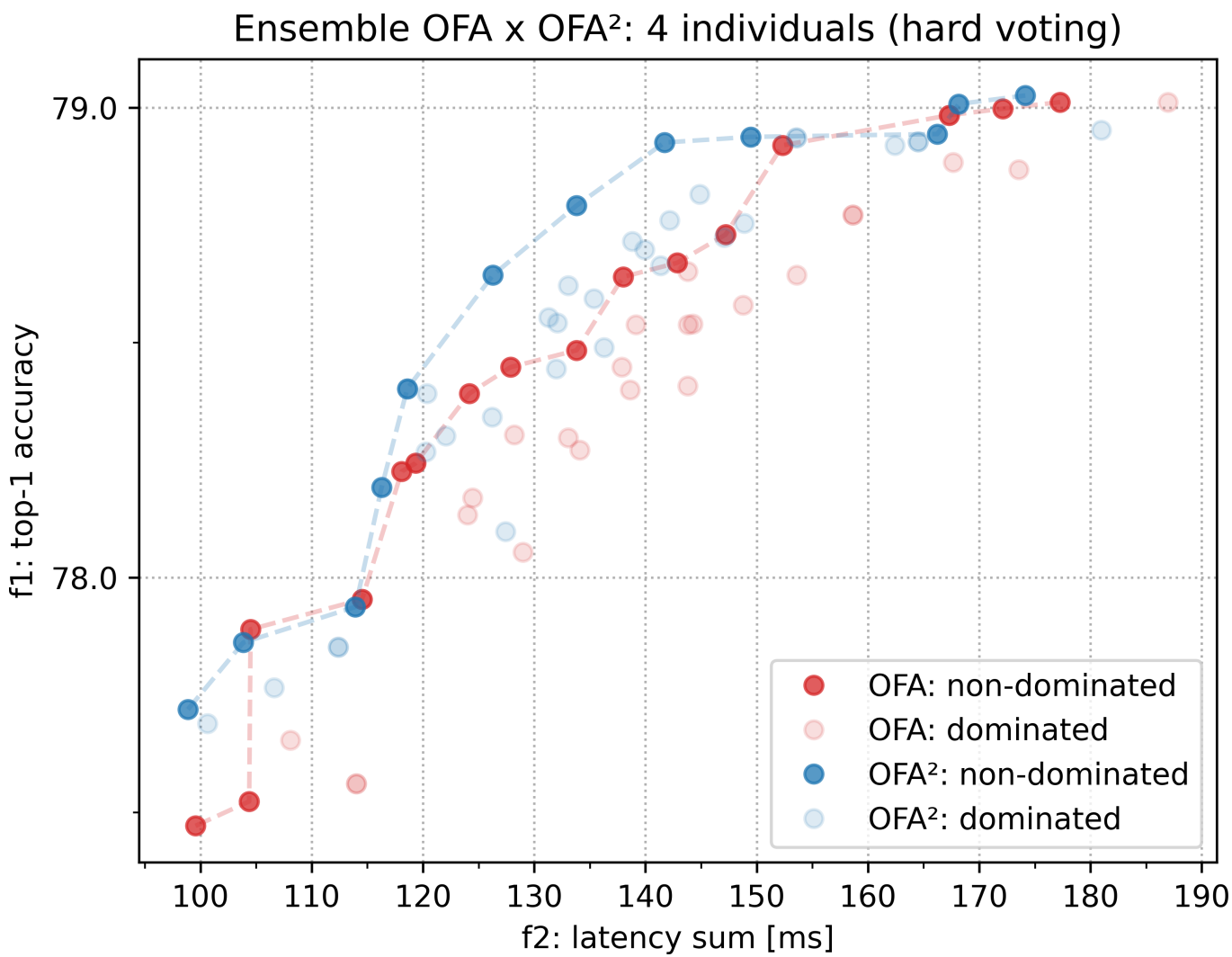
OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - [3 individuals]
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



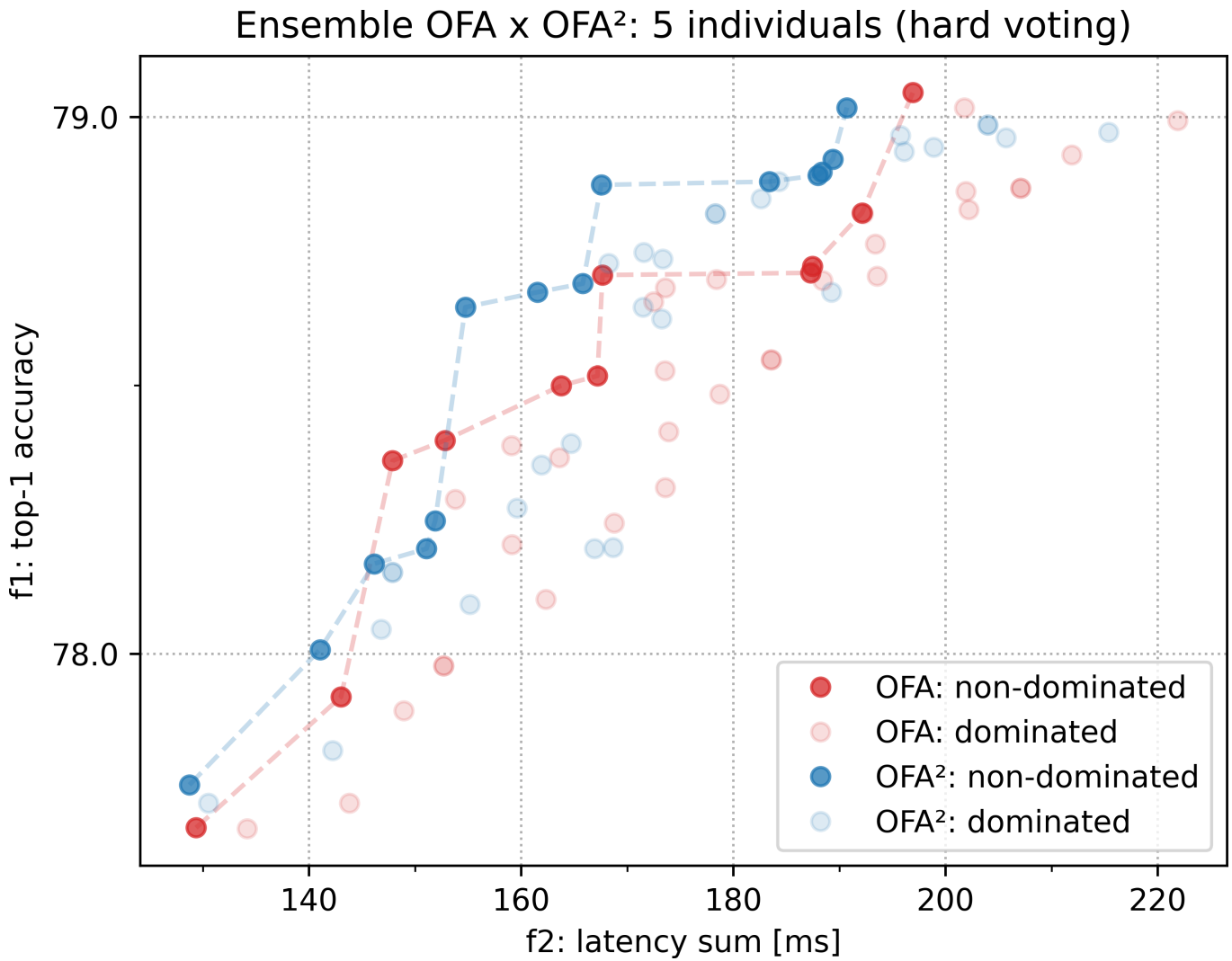
OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - [4 individuals]
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



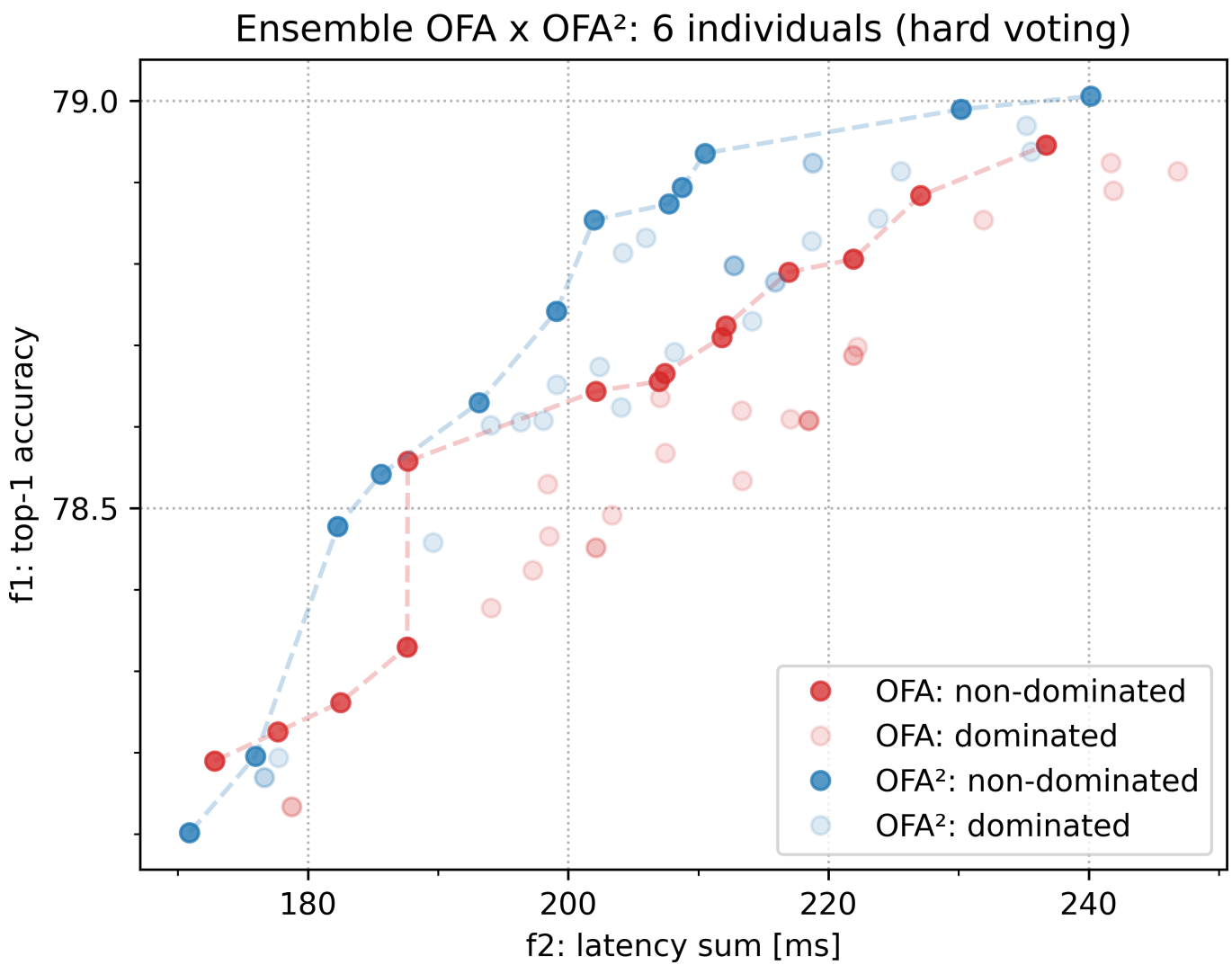
OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - [5 individuals]
  - 6 individuals
  - 7 individuals
  - 8 individuals



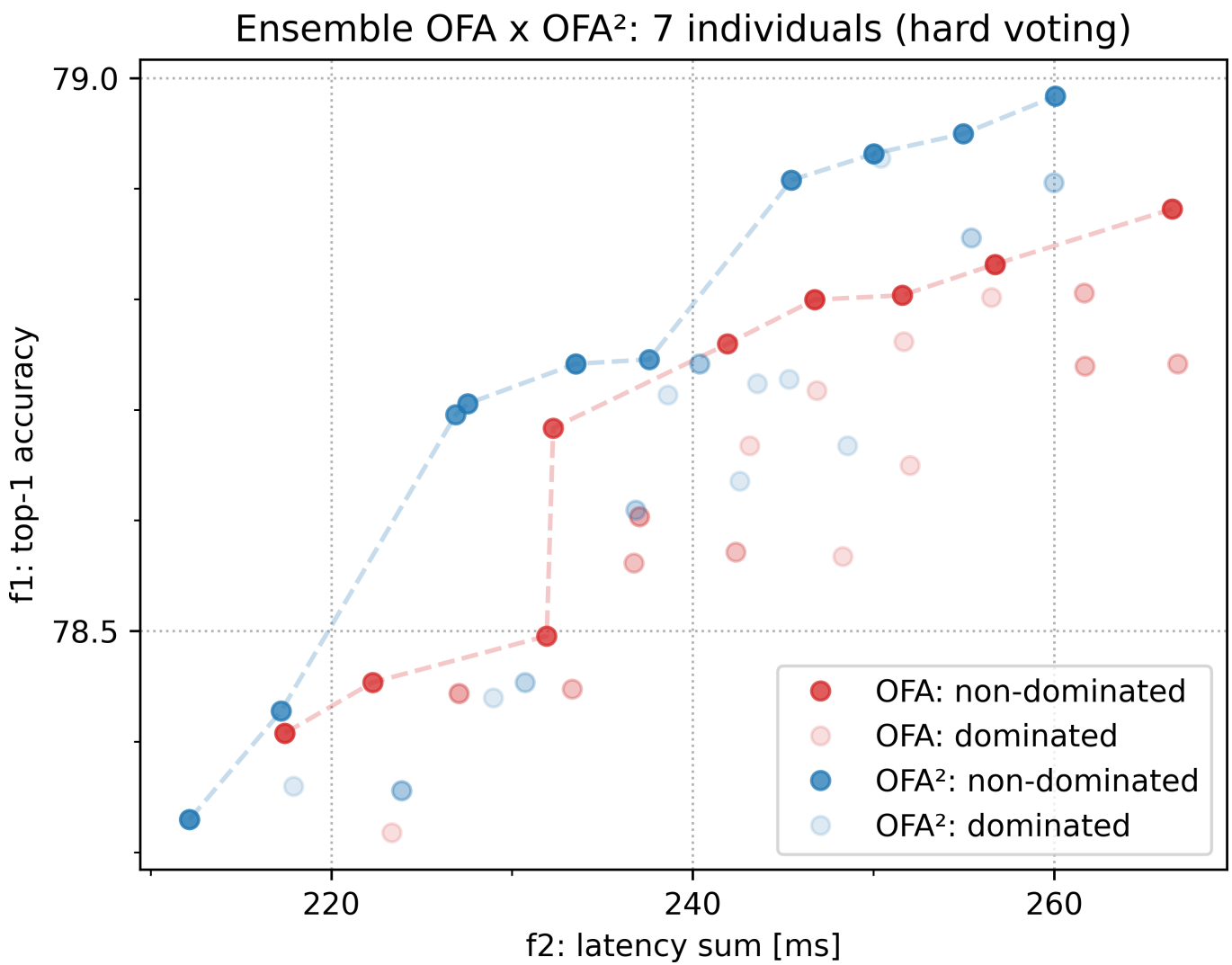
OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - [6 individuals]
  - 7 individuals
  - 8 individuals



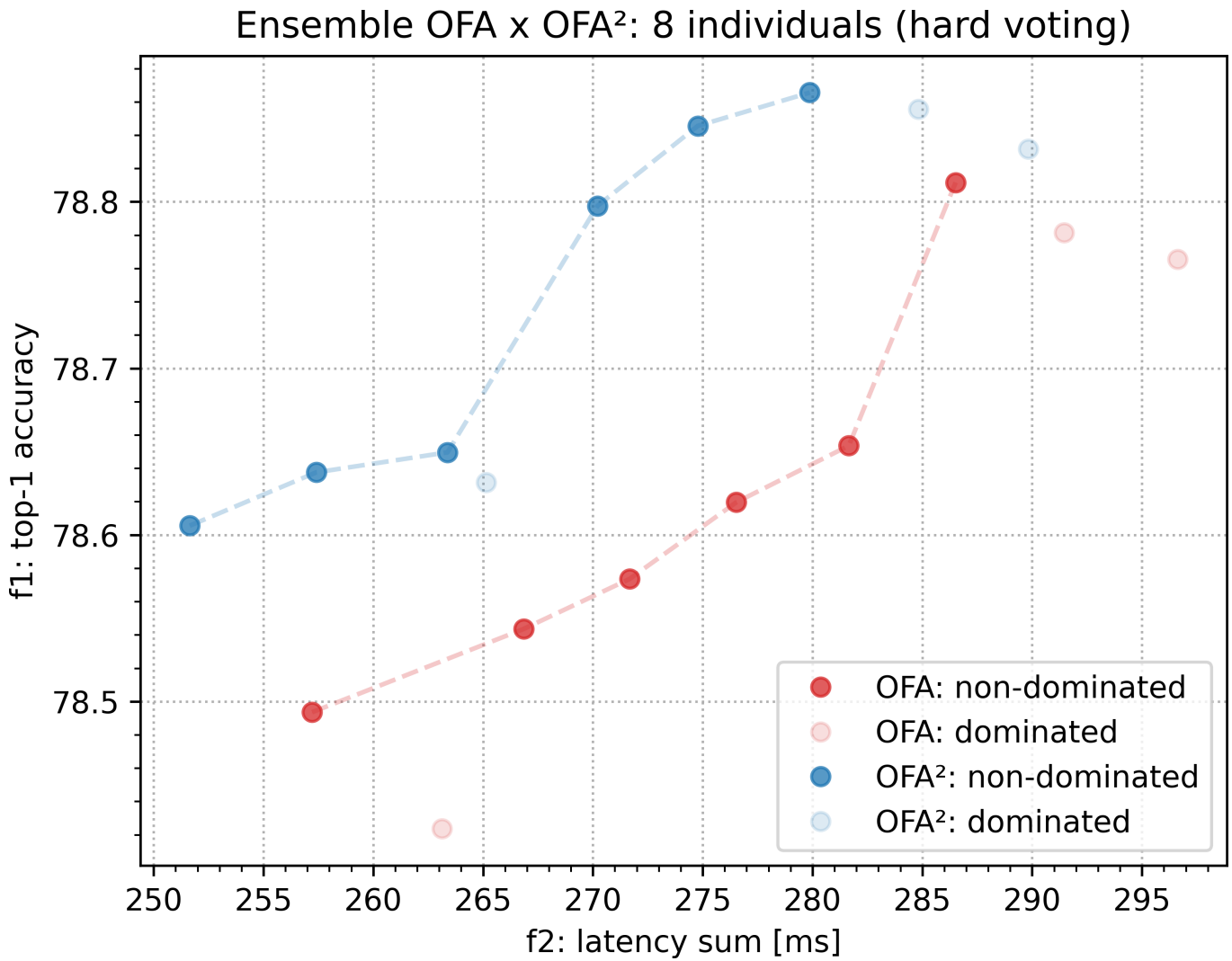
OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - [7 individuals]
  - 8 individuals



OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - [8 individuals]

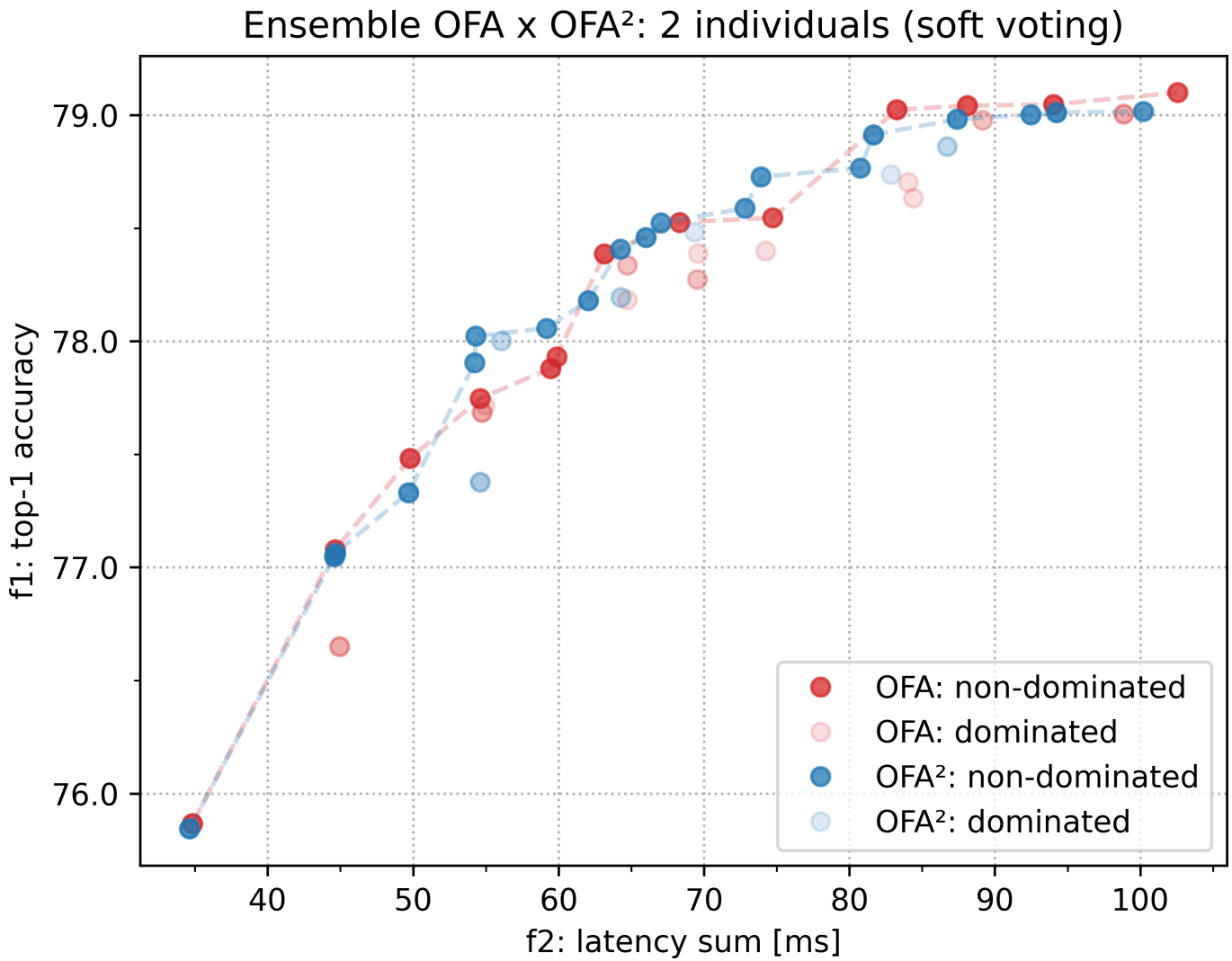


OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

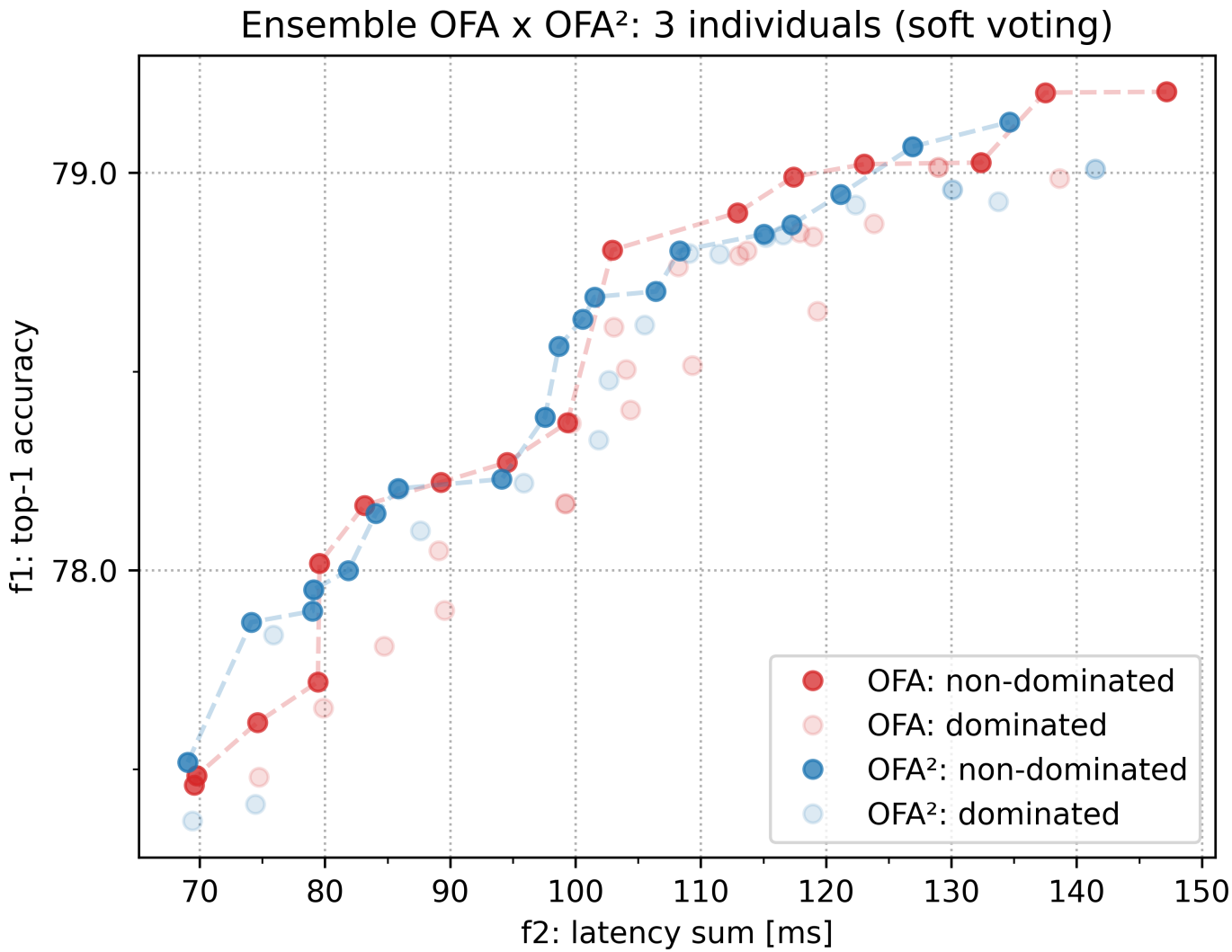
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



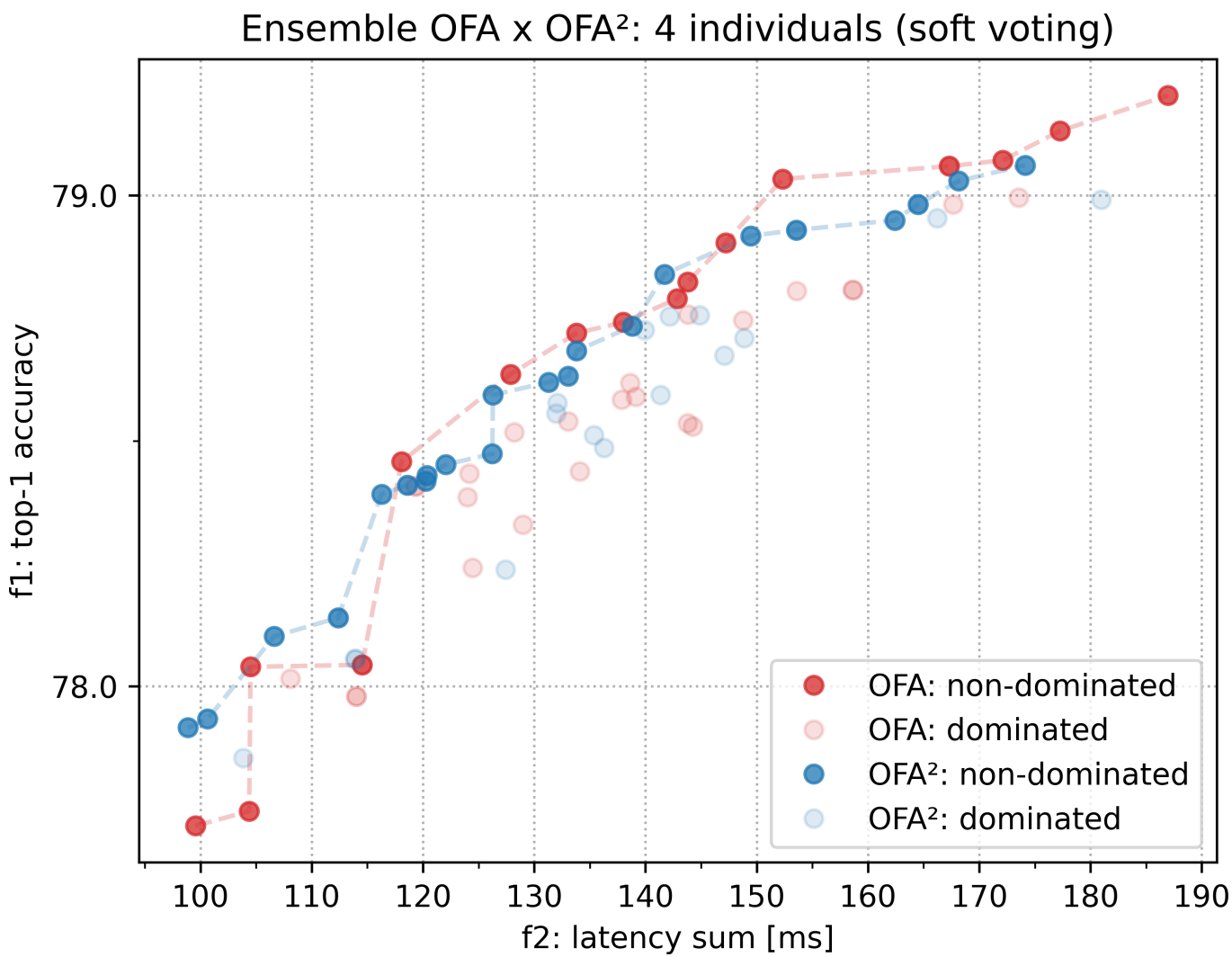
OFA² x OFA

- **Voting**
  - hard
  - [soft]
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - [3 individuals]
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



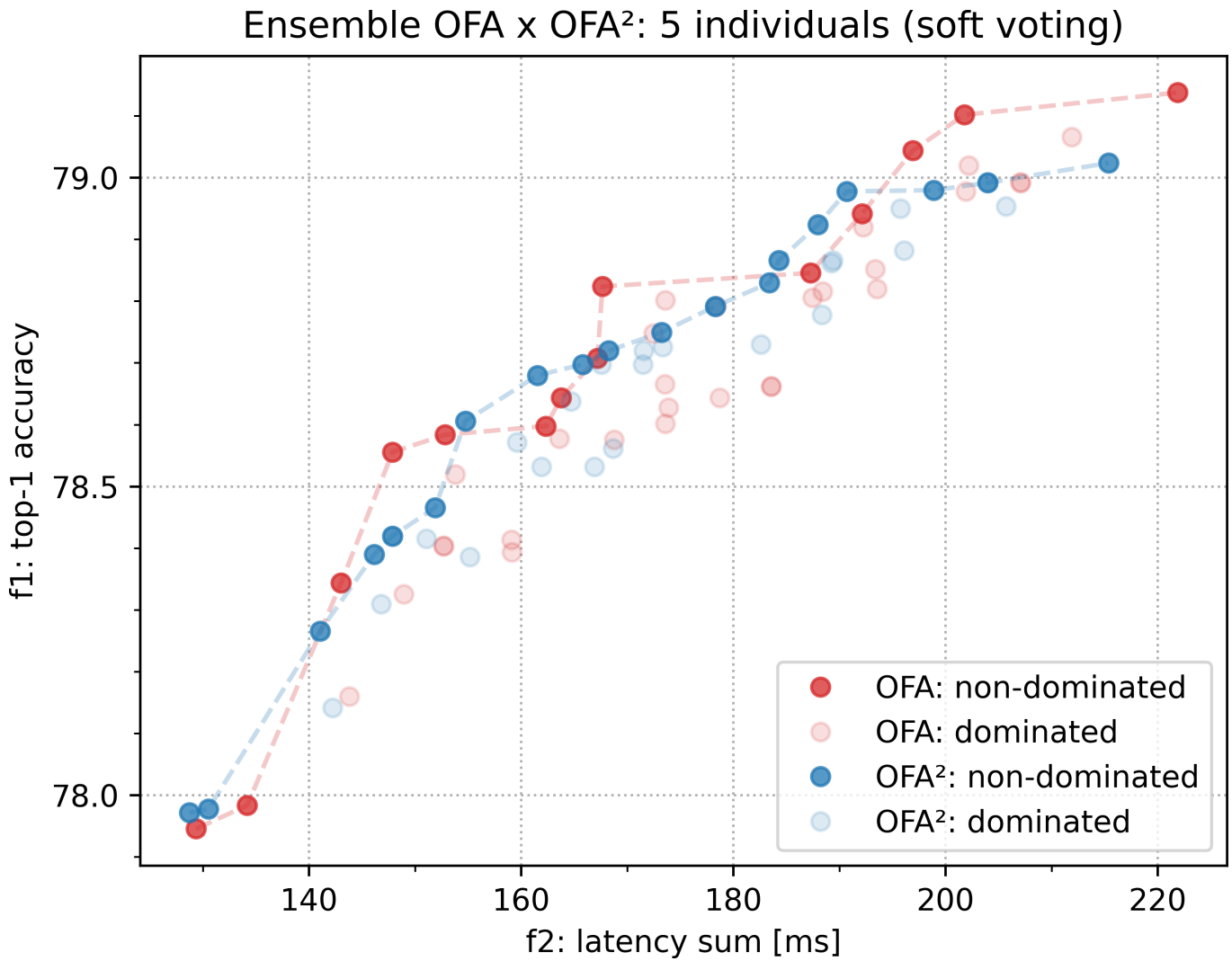
OFA² x OFA

- **Voting**
  - hard
  - [soft]
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - [4 individuals]
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



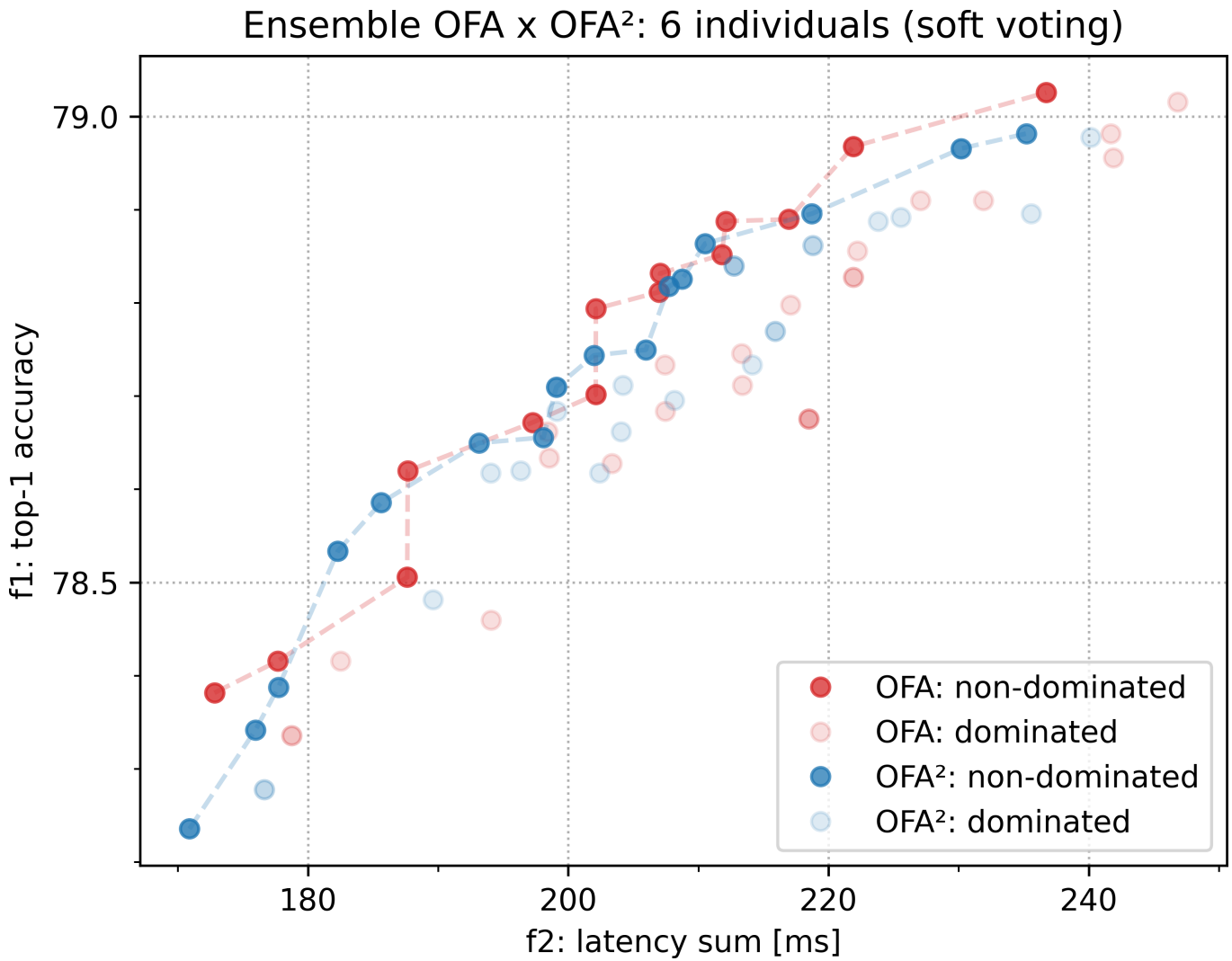
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



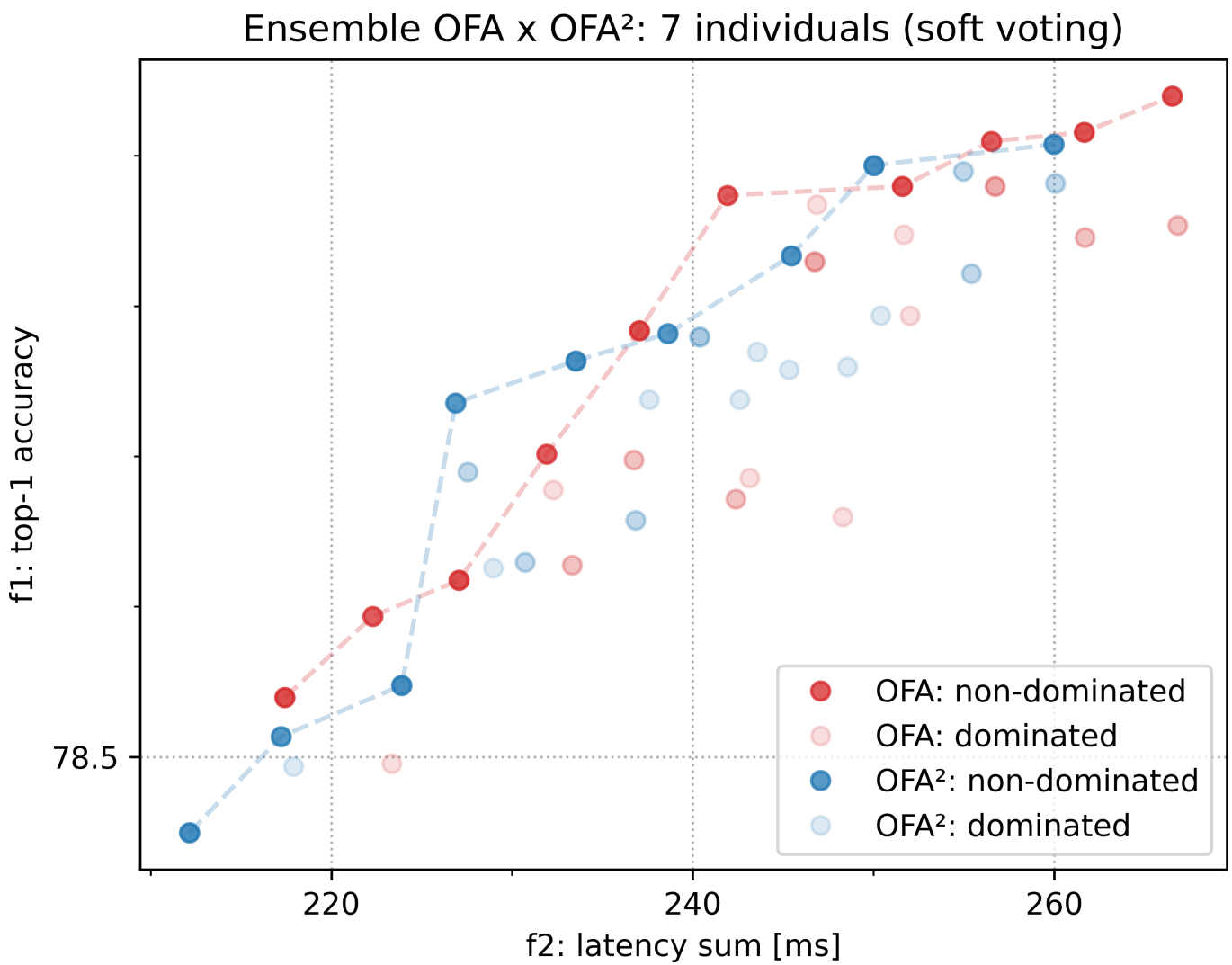
OFA² x OFA

- **Voting**
  - hard
  - [soft]
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - [6 individuals]
  - 7 individuals
  - 8 individuals



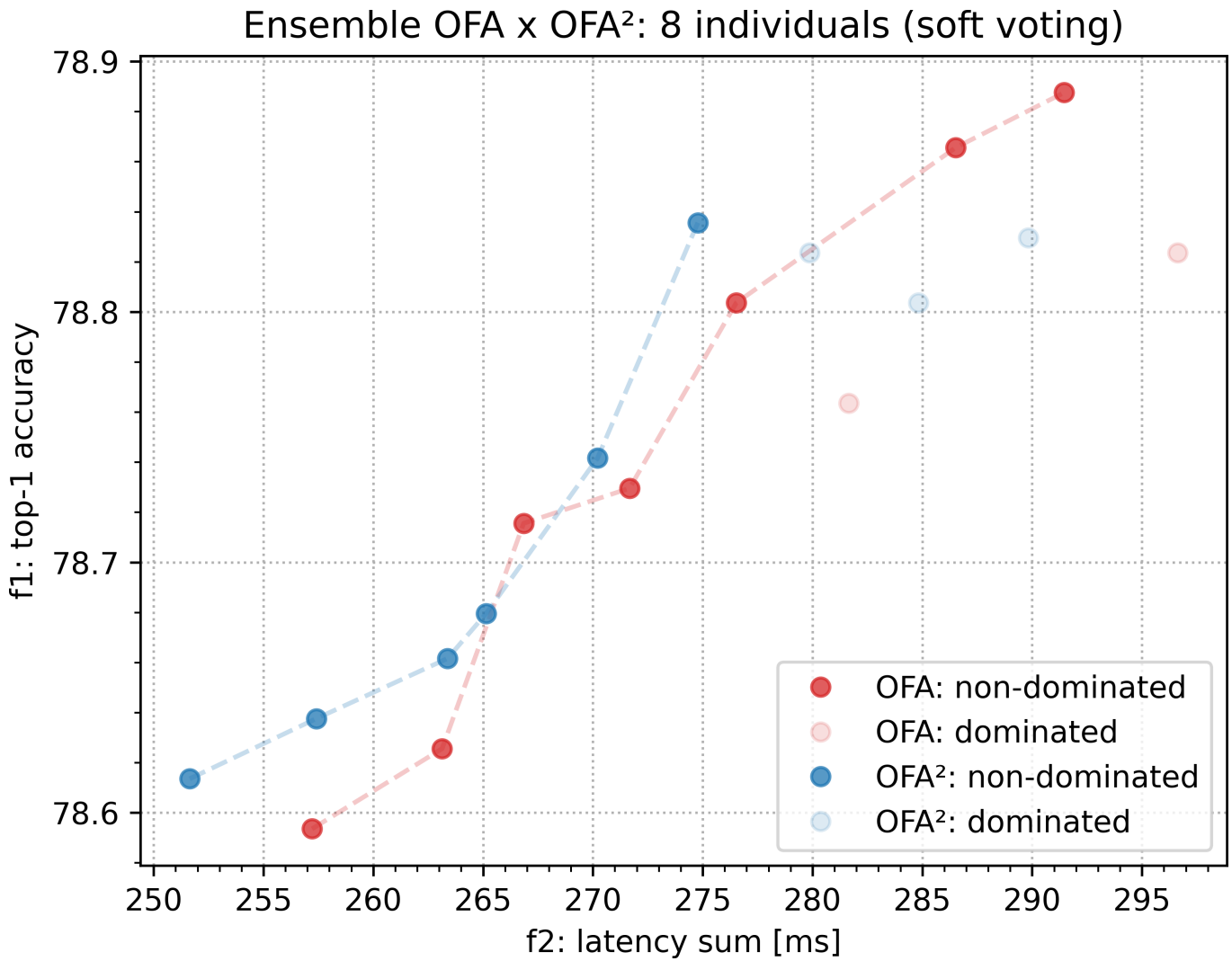
OFA² x OFA

- **Voting**
  - hard
  - [soft]
- **Latency**
  - [sum]
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - [7 individuals]
  - 8 individuals



OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

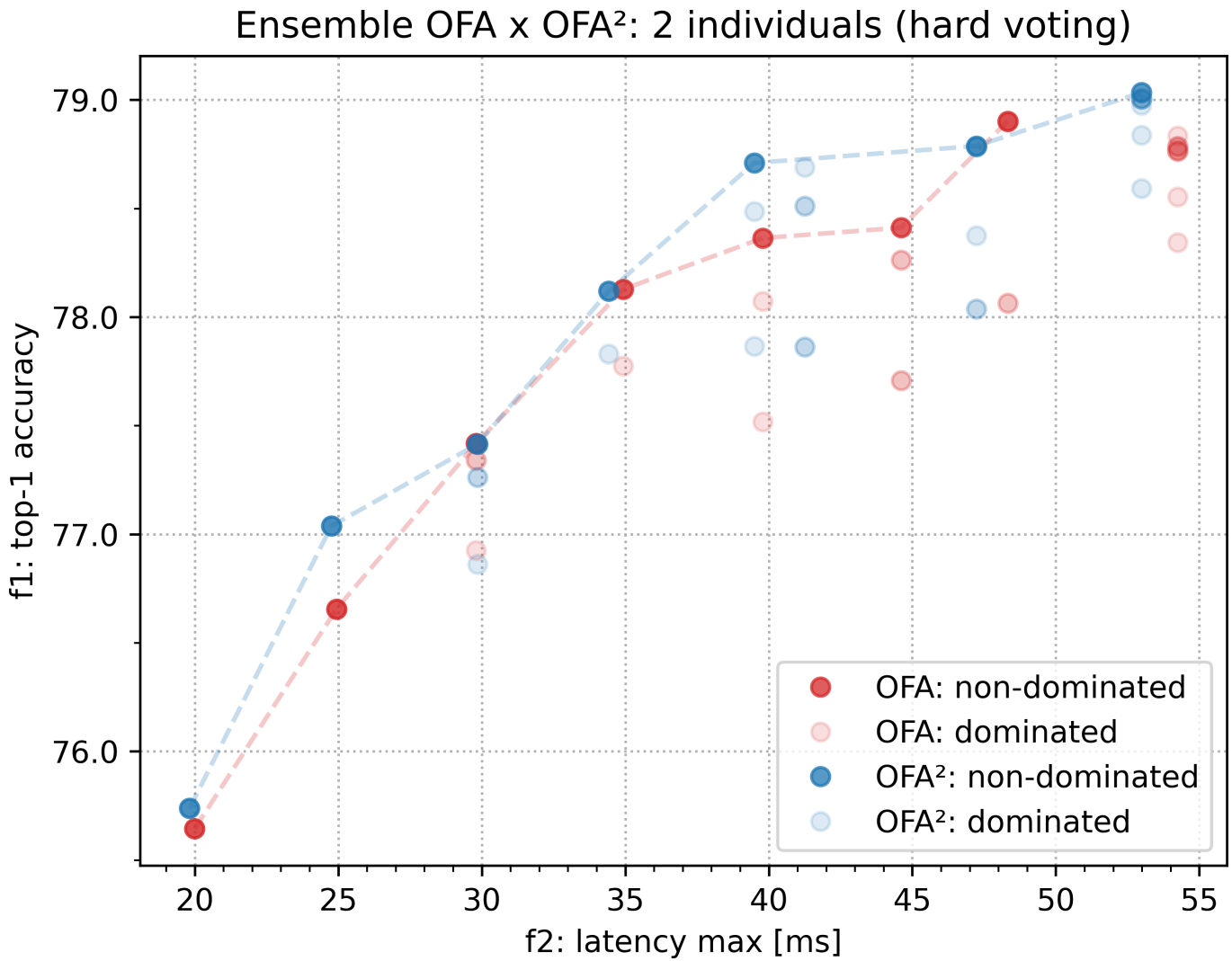


OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

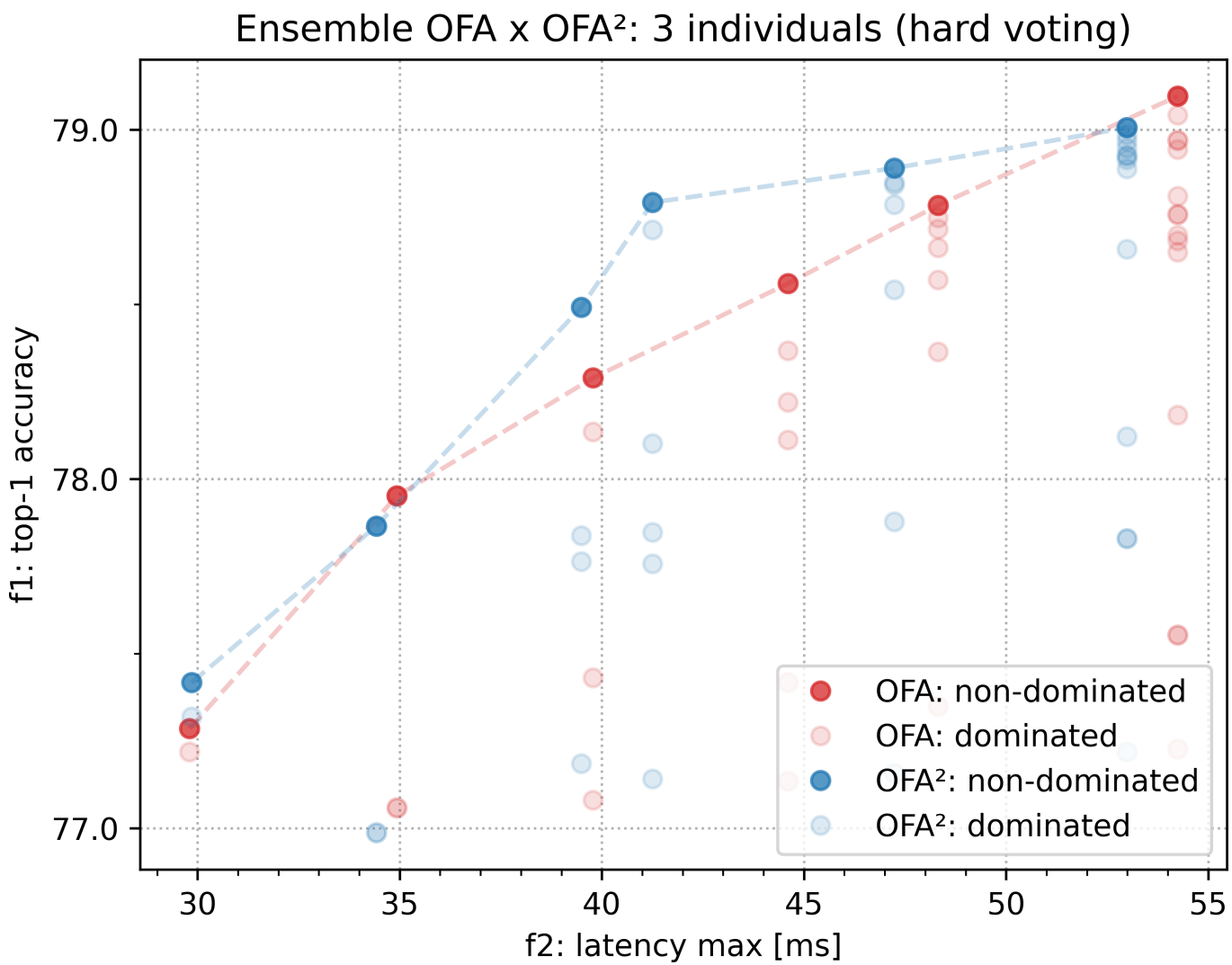
OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - sum
  - [max]
- **Individuals**
  - [2 individuals]
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



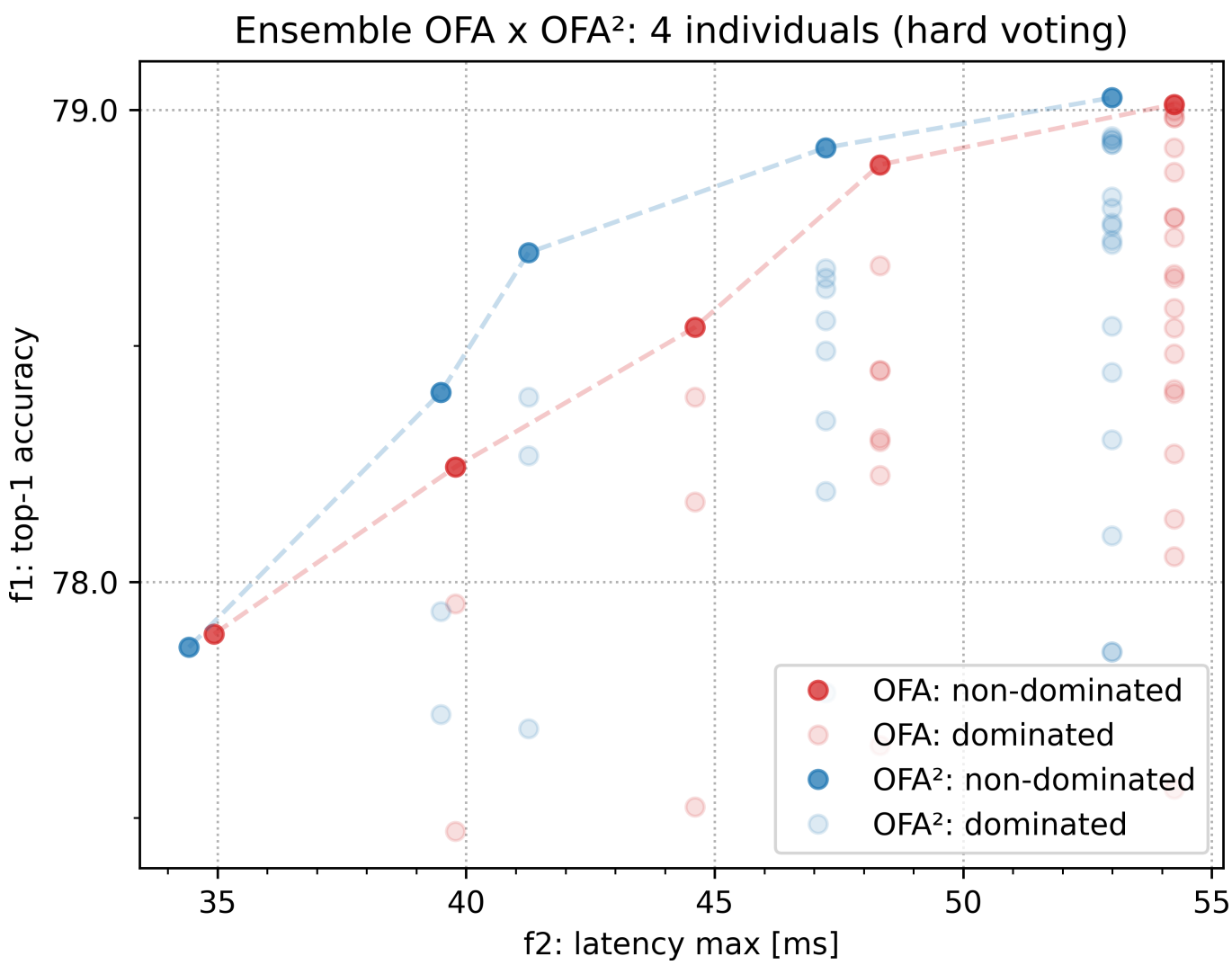
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



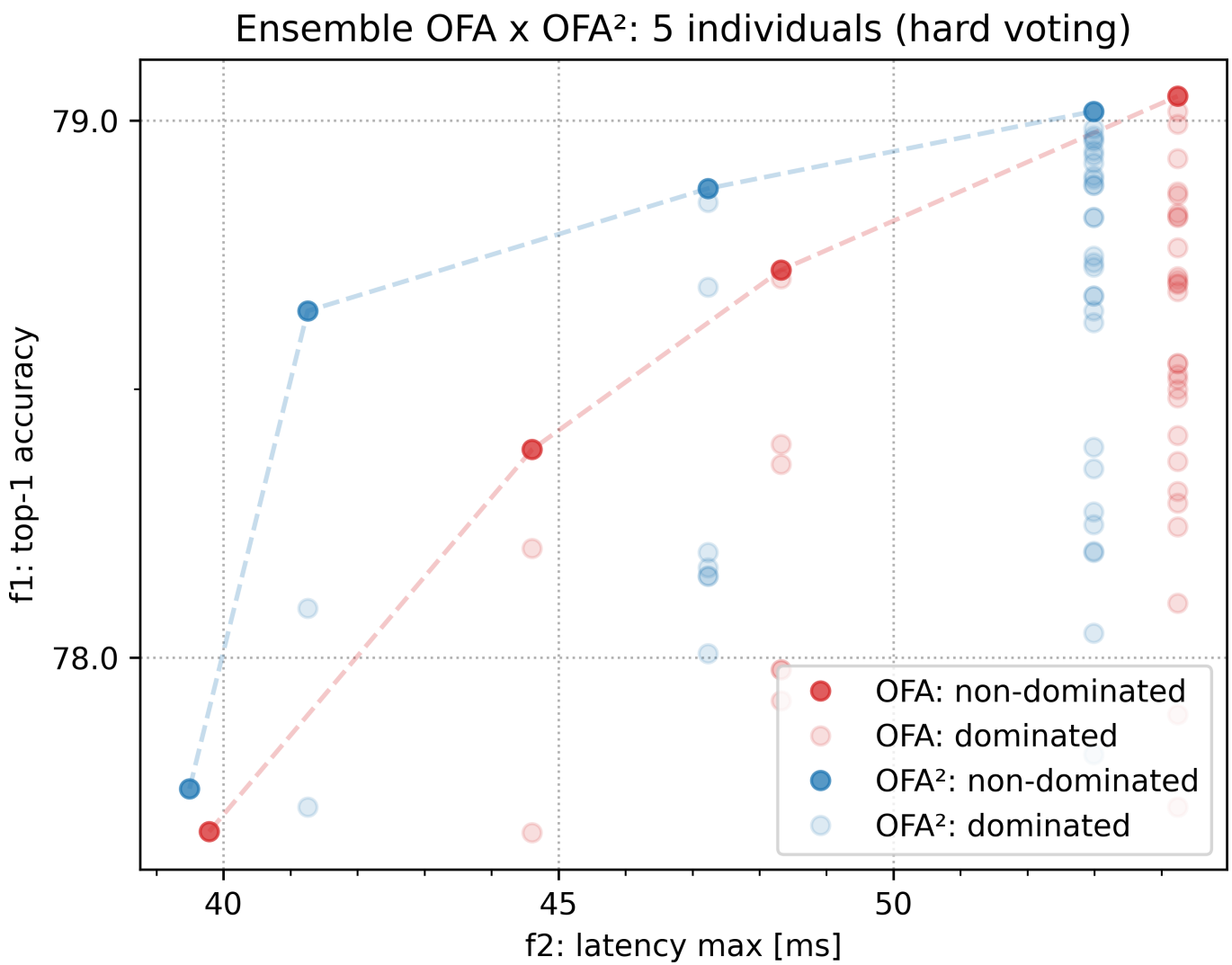
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



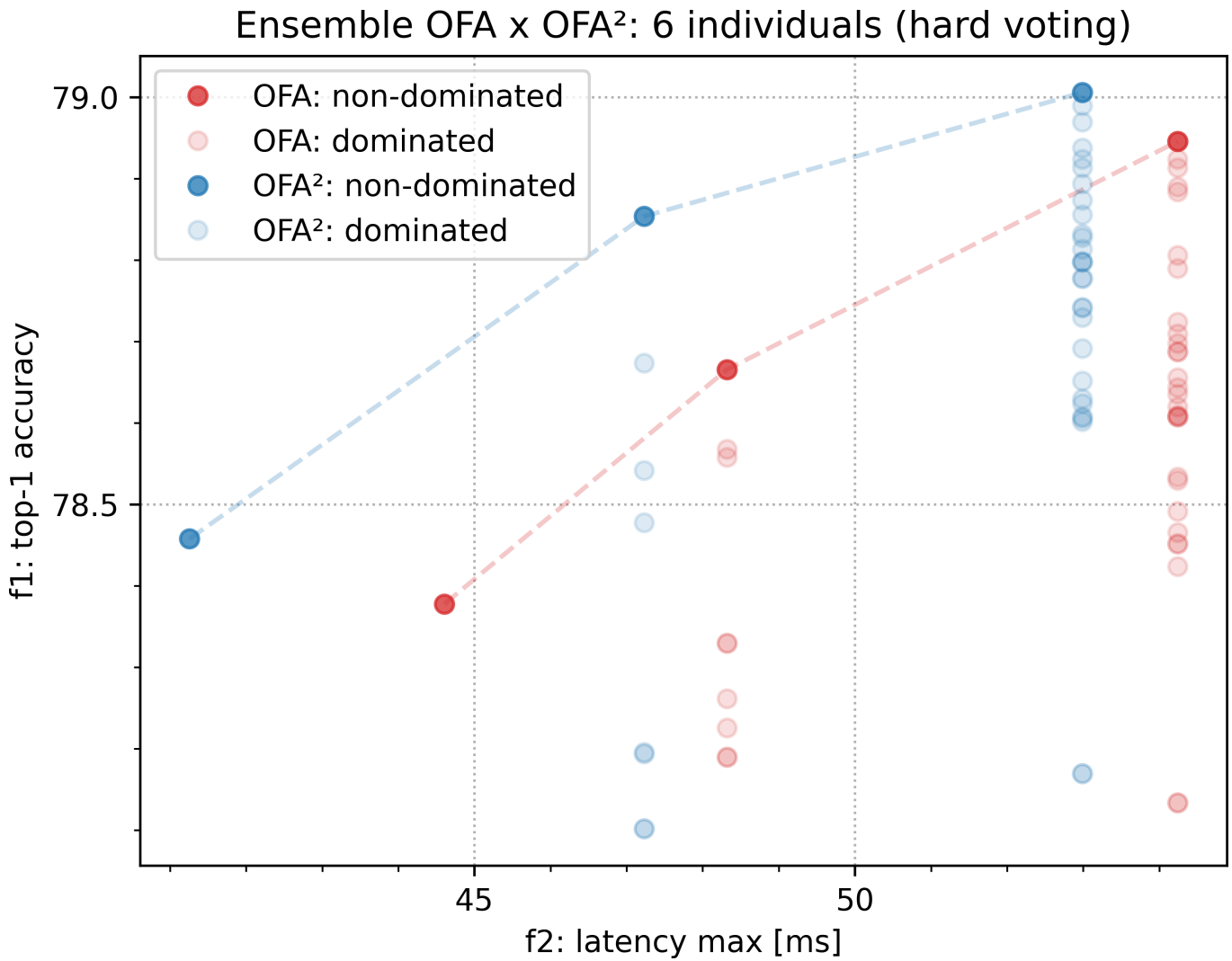
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



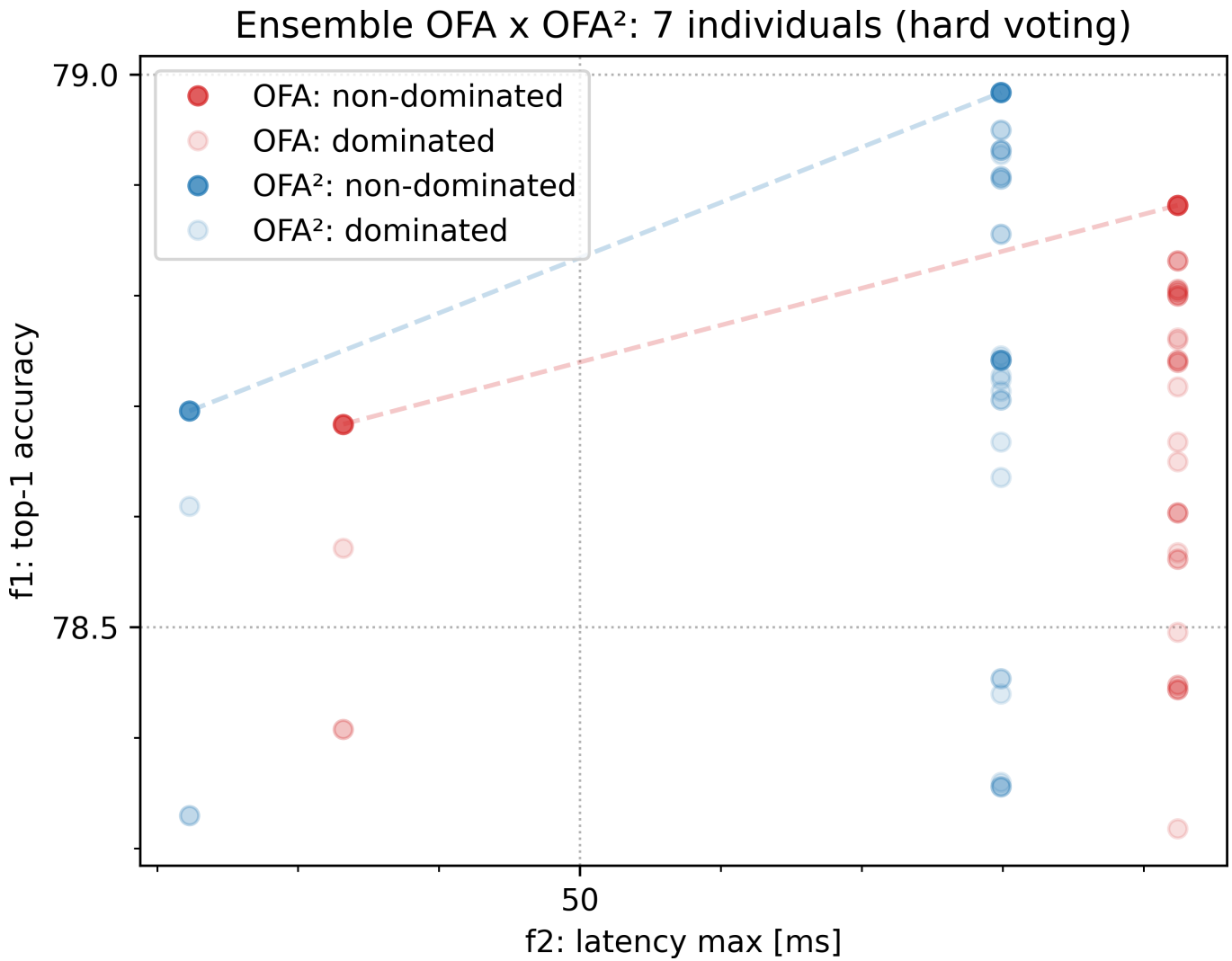
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



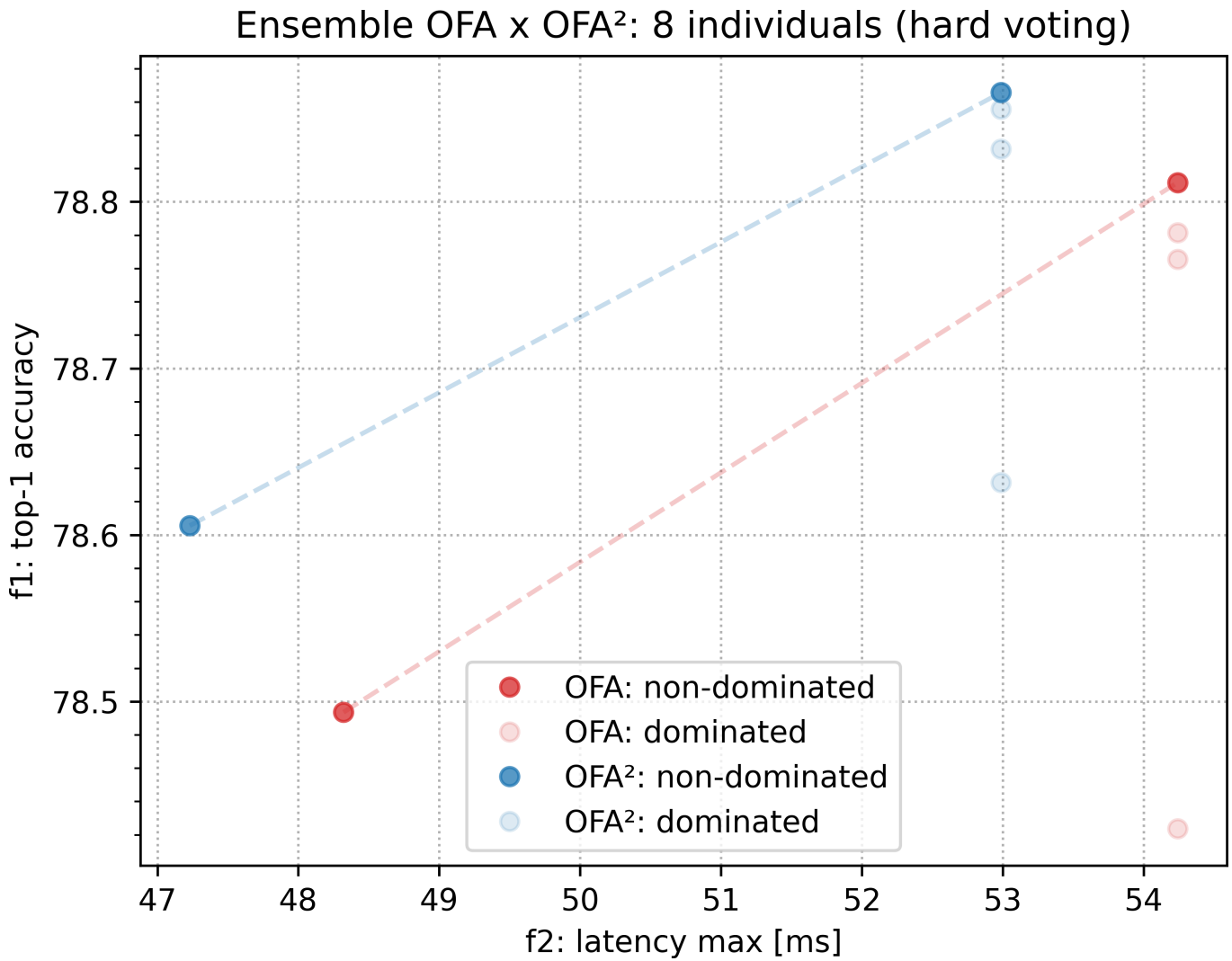
OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - [7 individuals]
  - 8 individuals



OFA² x OFA

- **Voting**
  - [hard]
  - soft
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - [8 individuals]

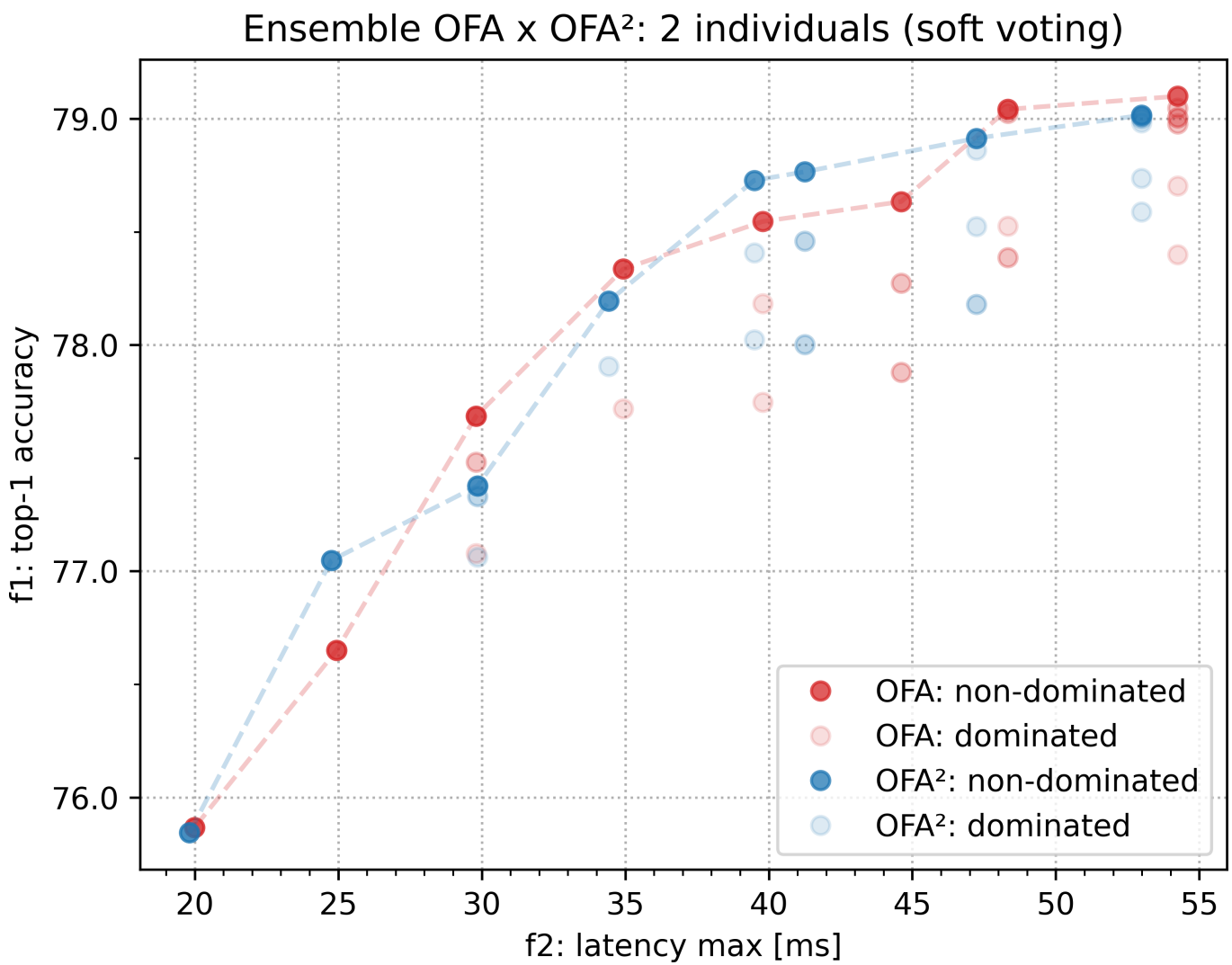


OFA² x OFA

- **Voting**
  - hard
  - [soft]
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

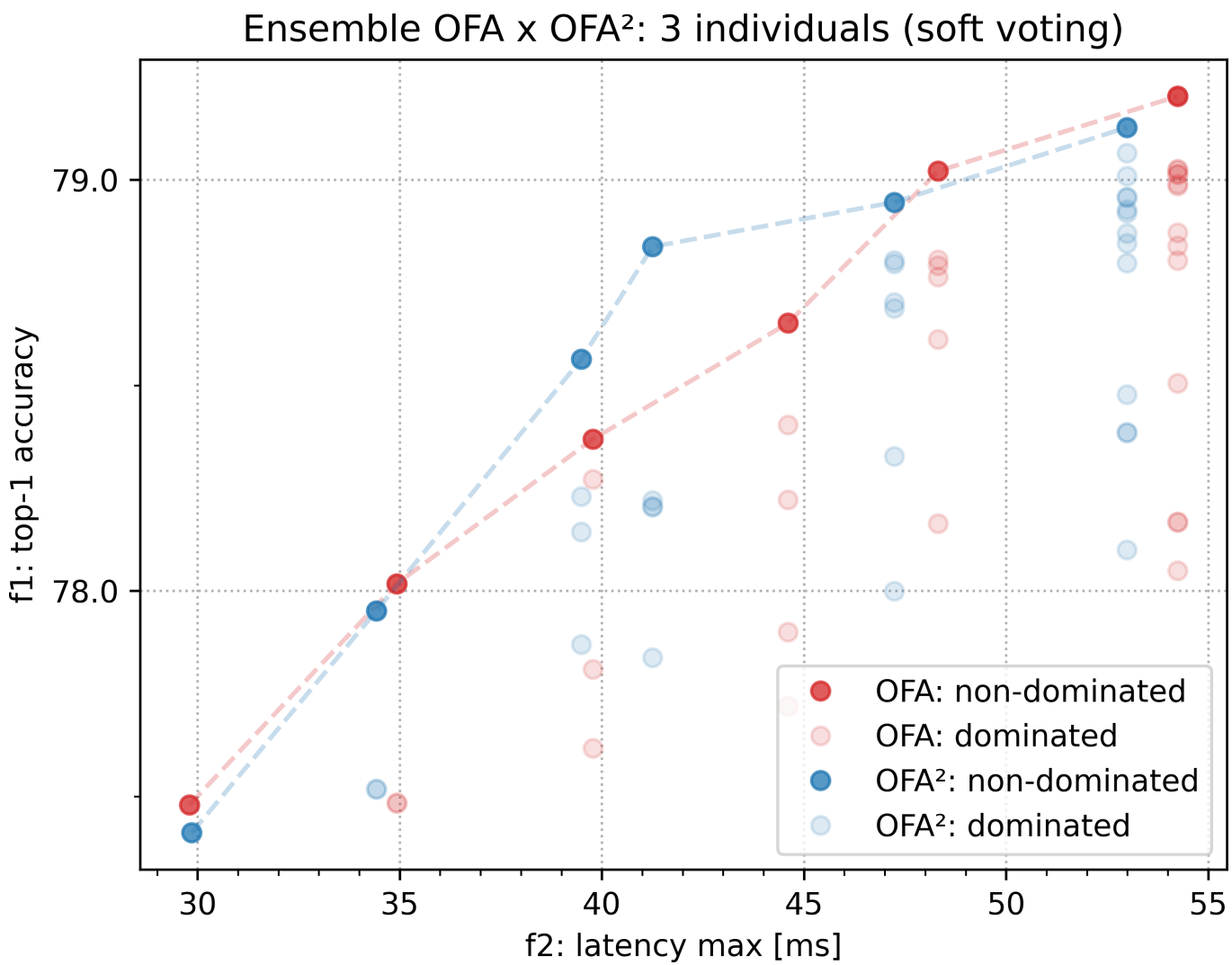
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



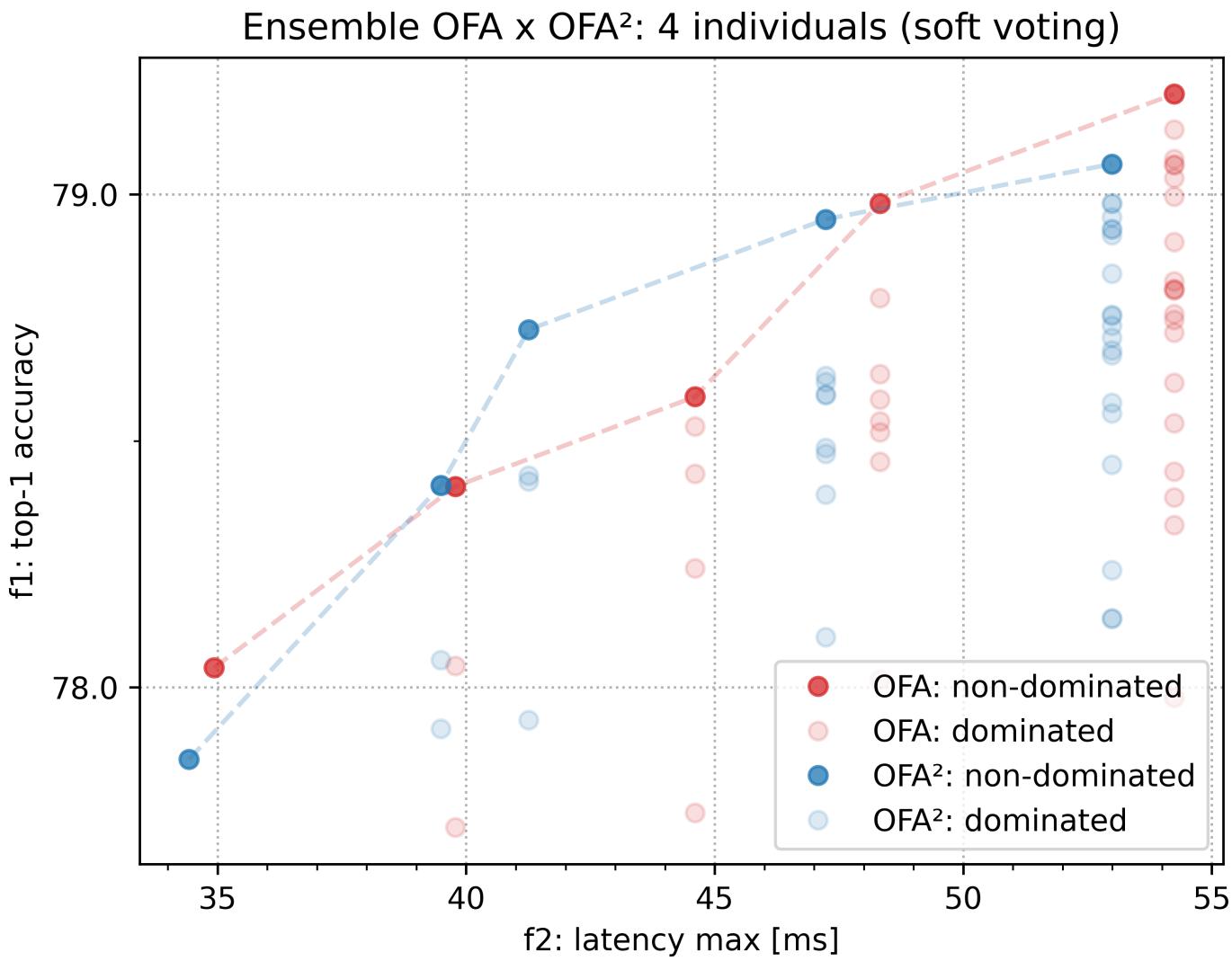
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



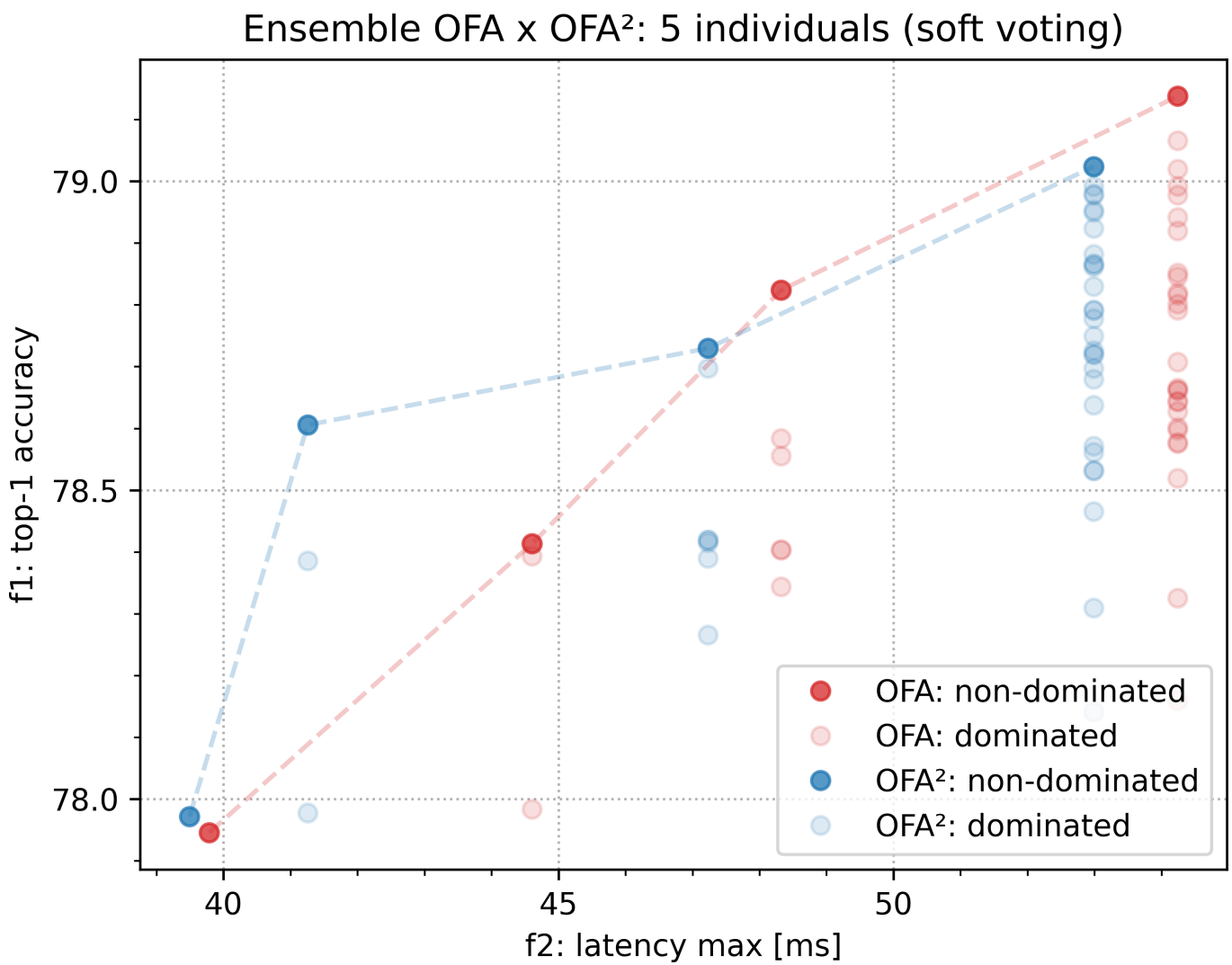
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



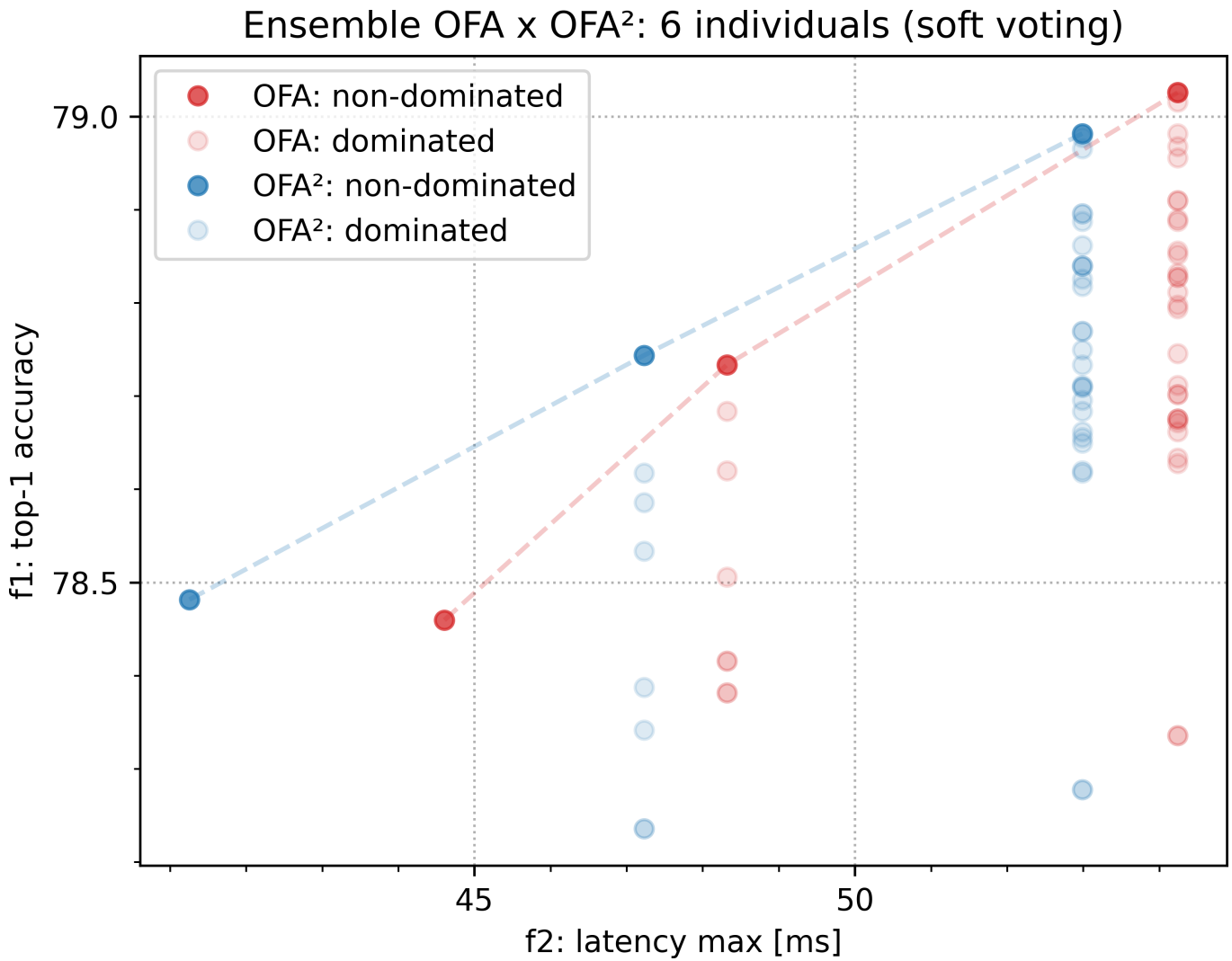
OFA² x OFA

- **Voting**
  - hard
  - [soft]
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - [5 individuals]
  - 6 individuals
  - 7 individuals
  - 8 individuals



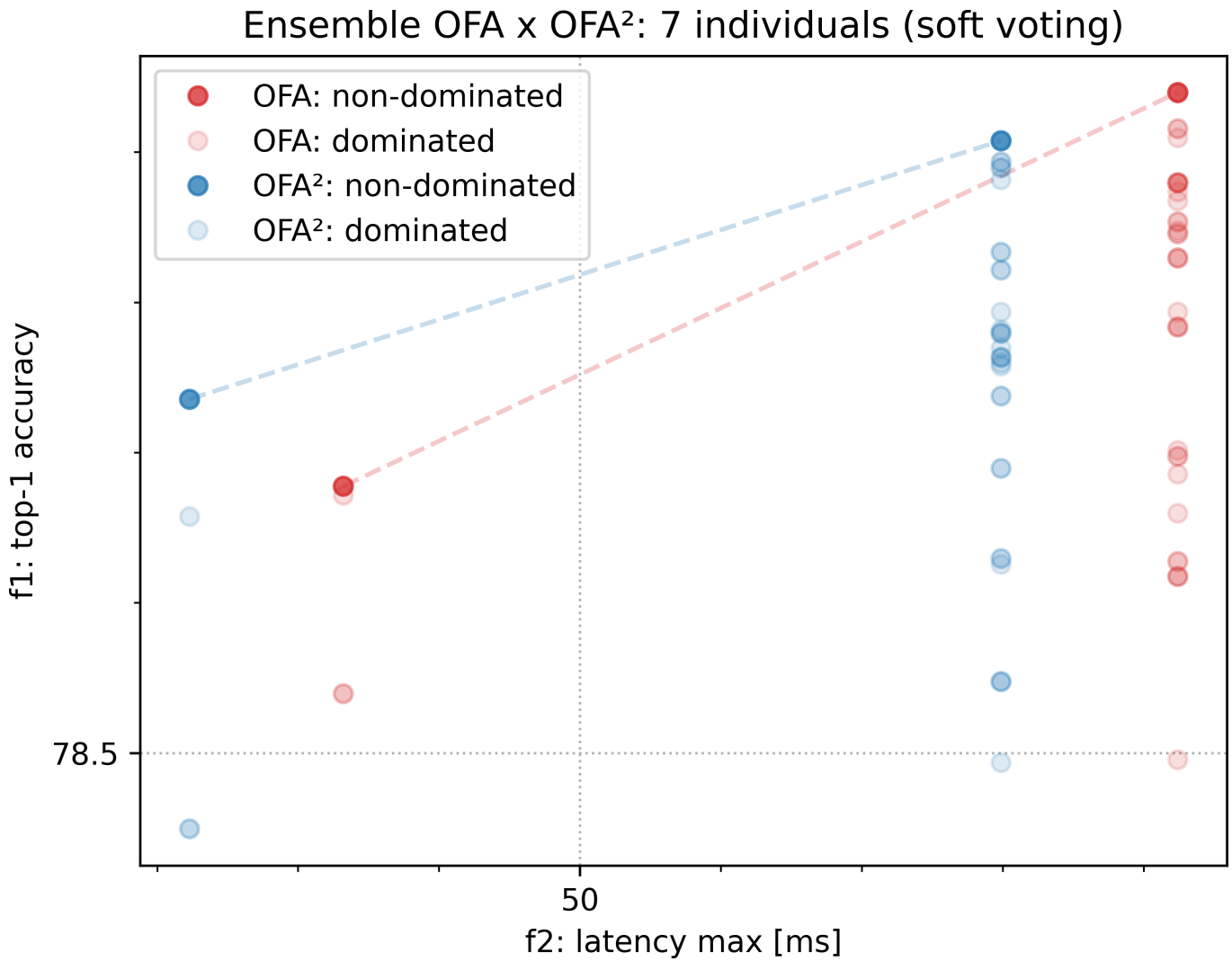
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals



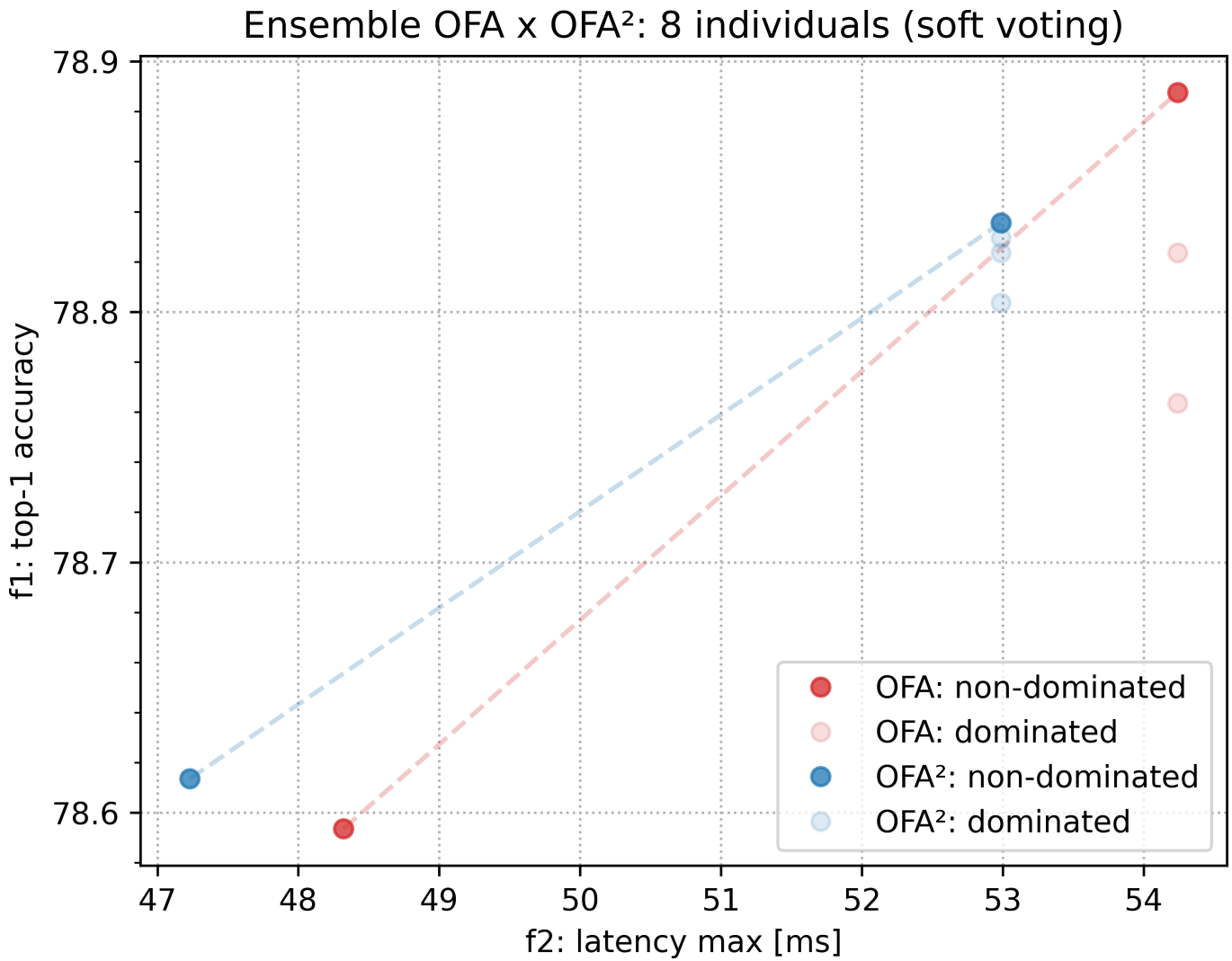
OFA² x OFA

- **Voting**
  - hard
  - [soft]
- **Latency**
  - sum
  - [max]
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - [7 individuals]
  - 8 individuals



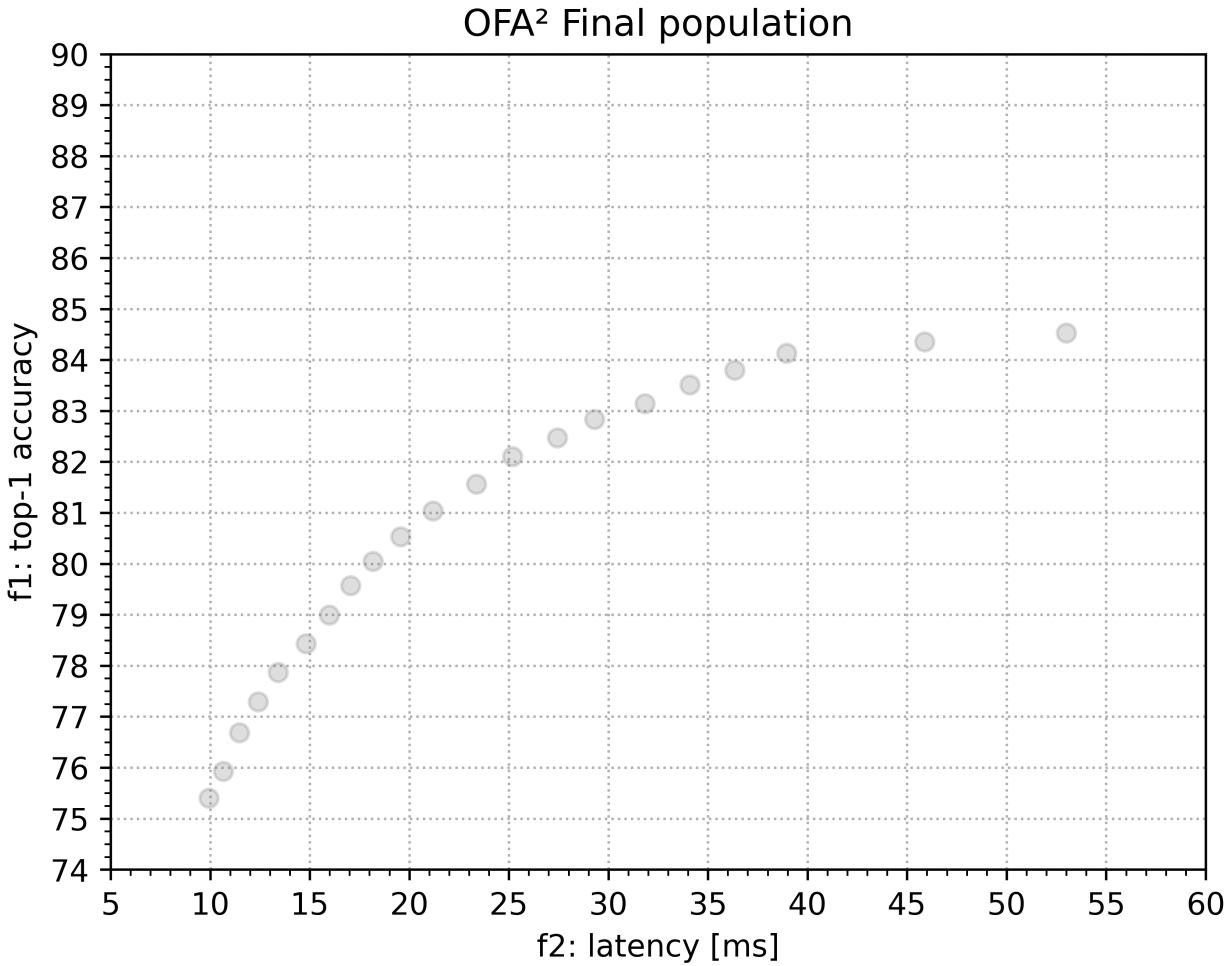
OFA² x OFA

- **Voting**
  - hard
  - soft
- **Latency**
  - sum
  - max
- **Individuals**
  - 2 individuals
  - 3 individuals
  - 4 individuals
  - 5 individuals
  - 6 individuals
  - 7 individuals
  - 8 individuals

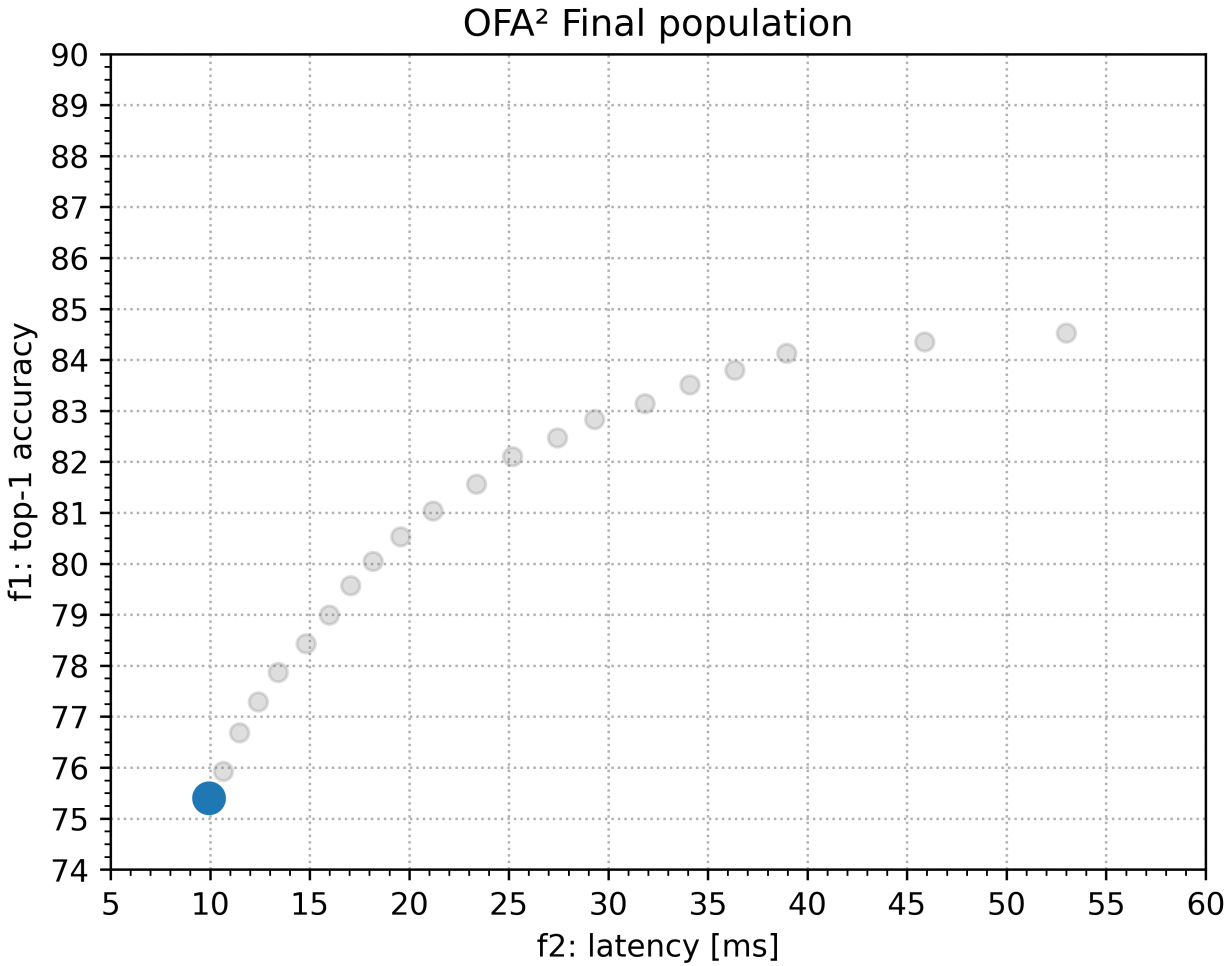


Agenda

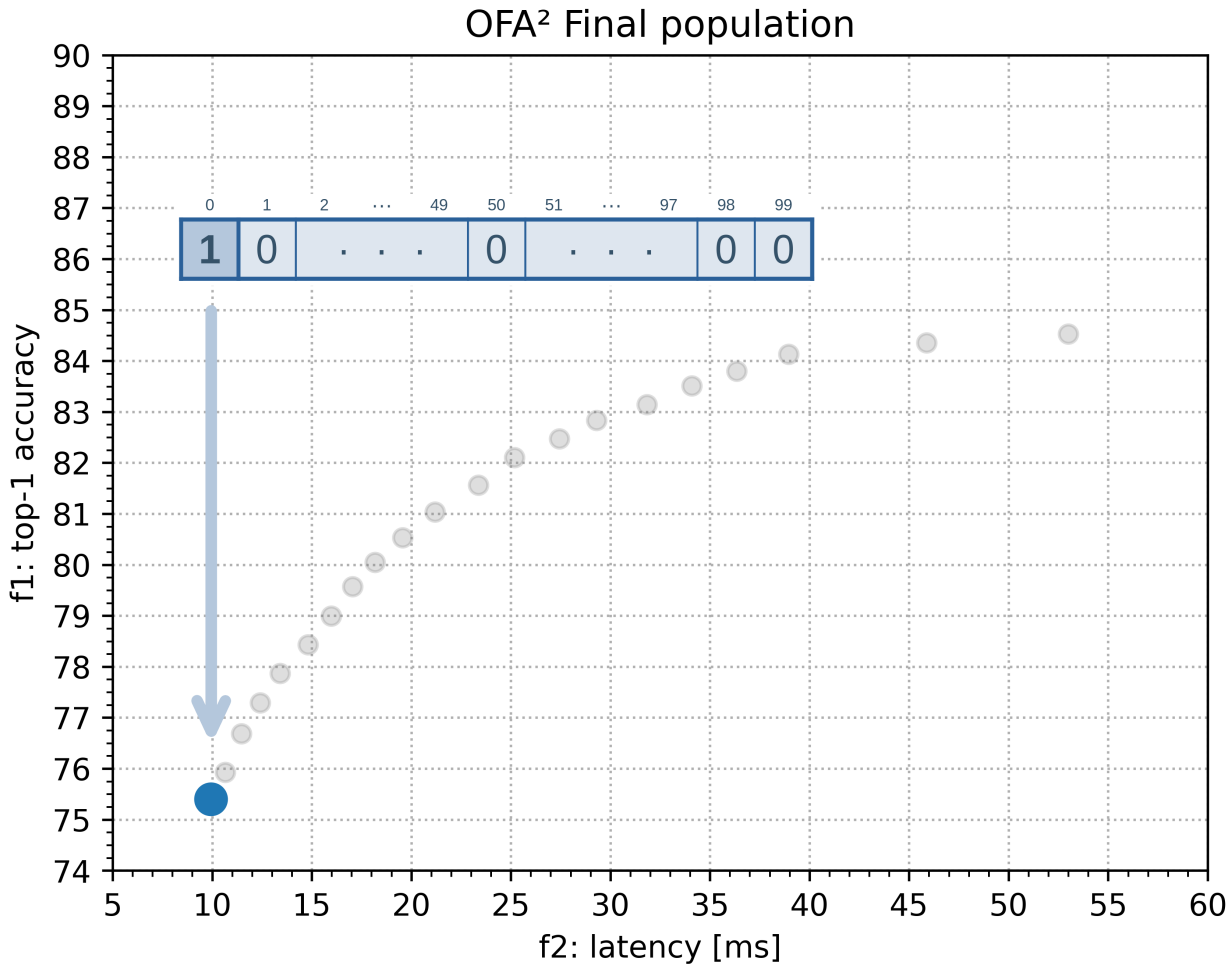
Encoding


OFA² Final population
f1: top-1 accuracy
f2: latency [ms]
90
89
88
87
86
85
84
83
82
81
80
79
78
77
76
75
74
5
10
15
20
25
30
35
40
45
50
55
60

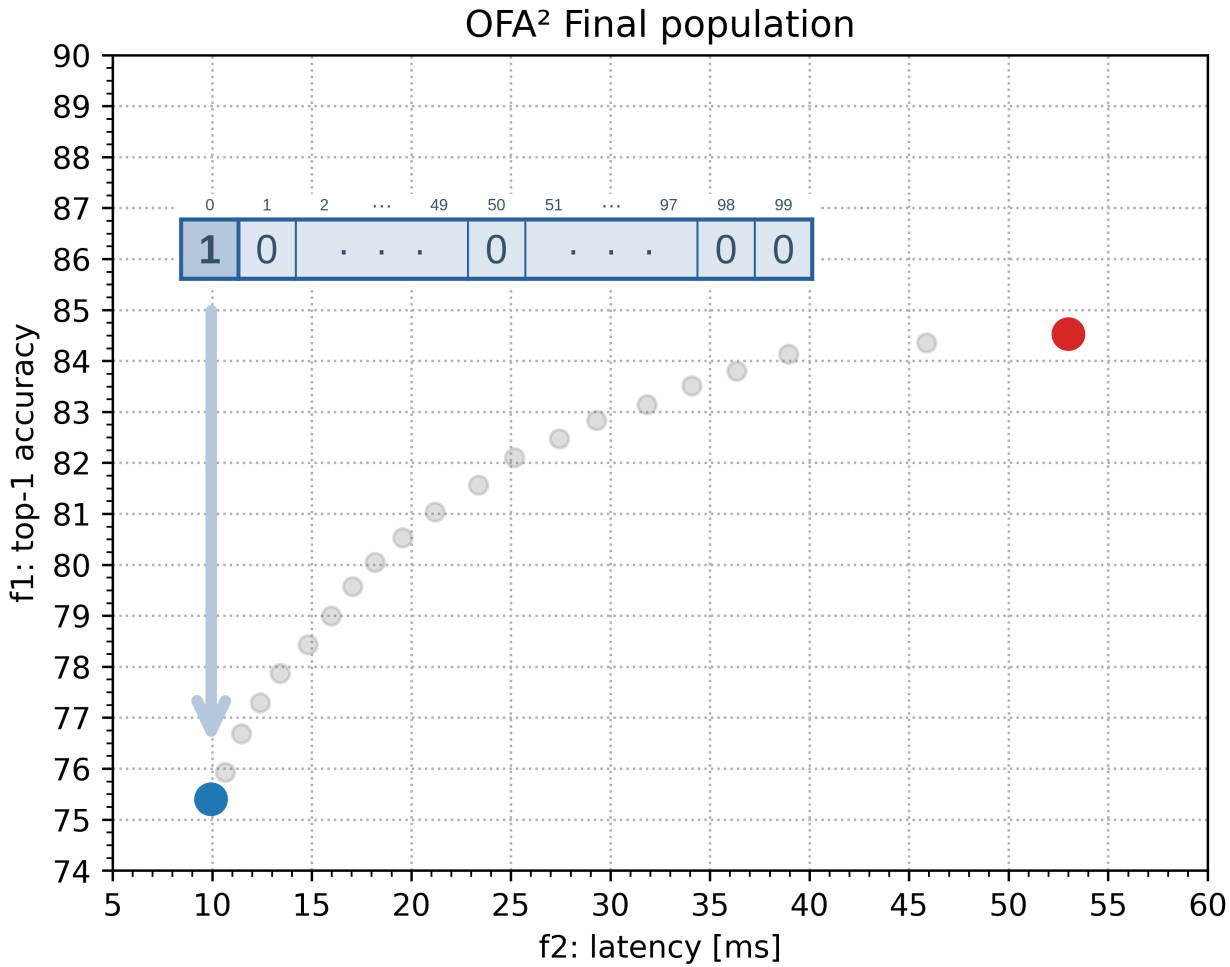
Encoding


OFA² Final population
f1: top-1 accuracy
f2: latency [ms]
90
89
88
87
86
85
84
83
82
81
80
79
78
77
76
75
74
5
10
15
20
25
30
35
40
45
50
55
60

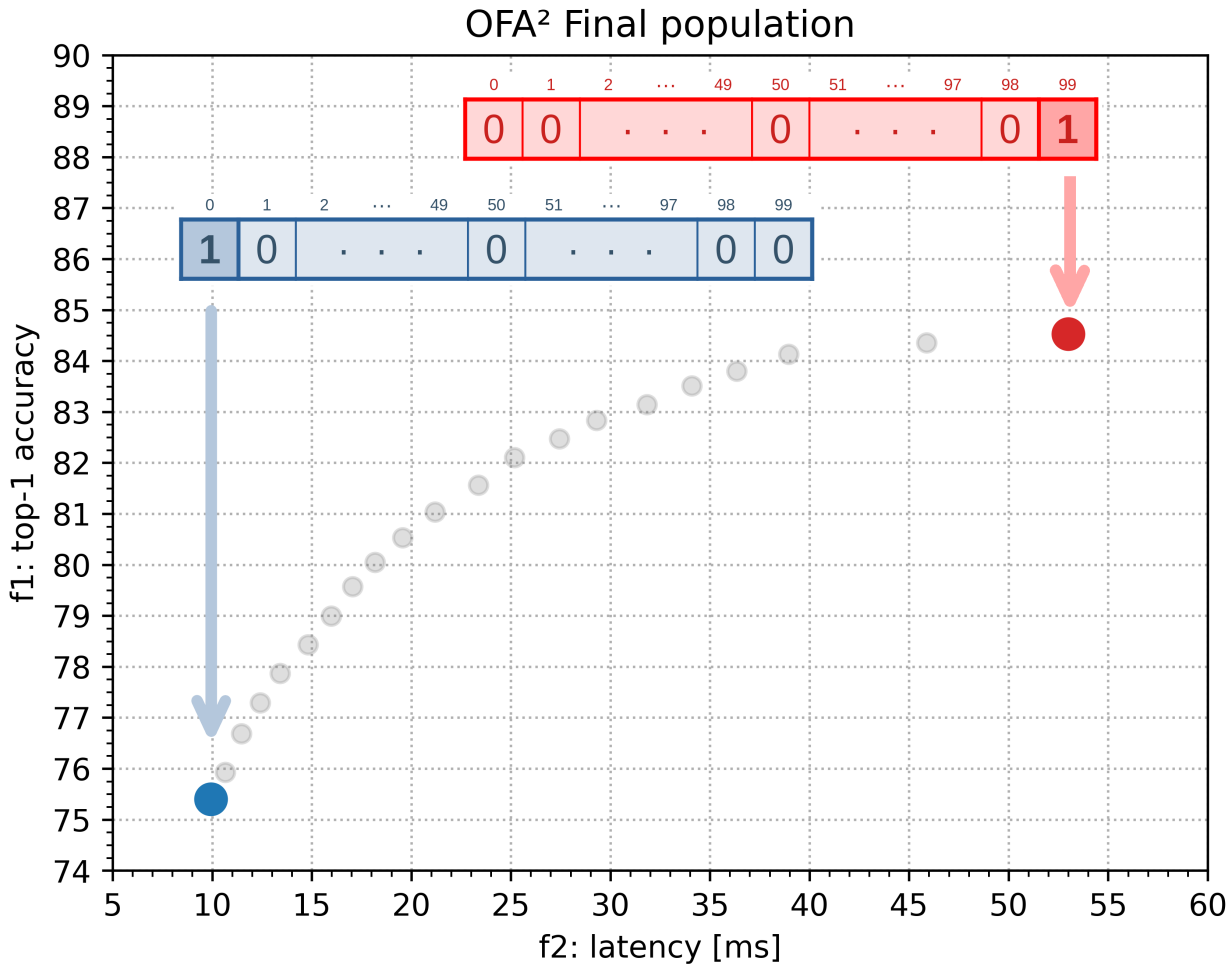
Encoding


OFA² Final population
f1: top-1 accuracy
f2: latency [ms]
0 1 2 ··· 49 50 51 ··· 97 98 99
1 0 · · · 0 · · · 0 0

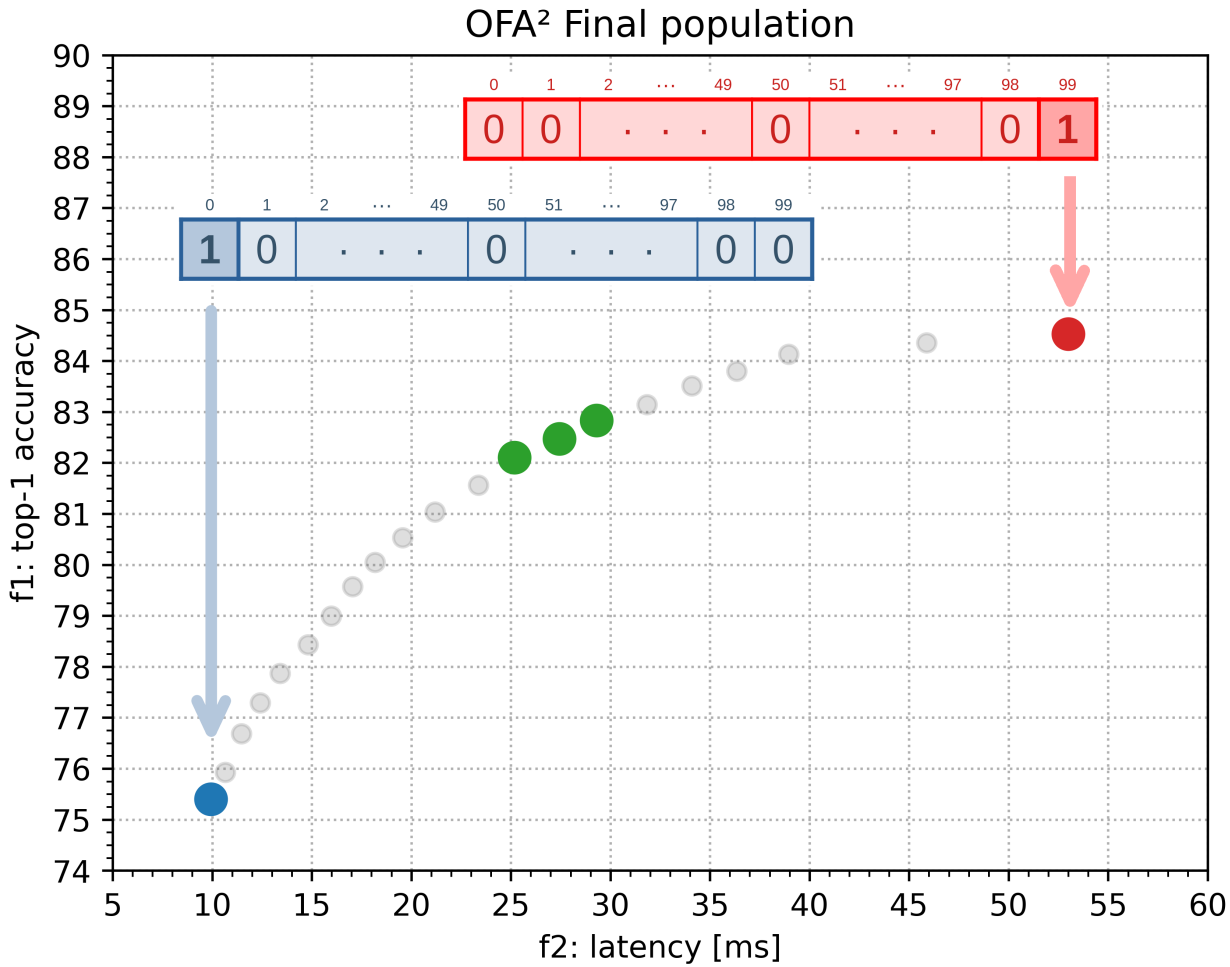
Encoding


OFA² Final population
f1: top-1 accuracy
f2: latency [ms]
0 1 2 ··· 49 50 51 ··· 97 98 99
1 0 · · · 0 · · · 0 0

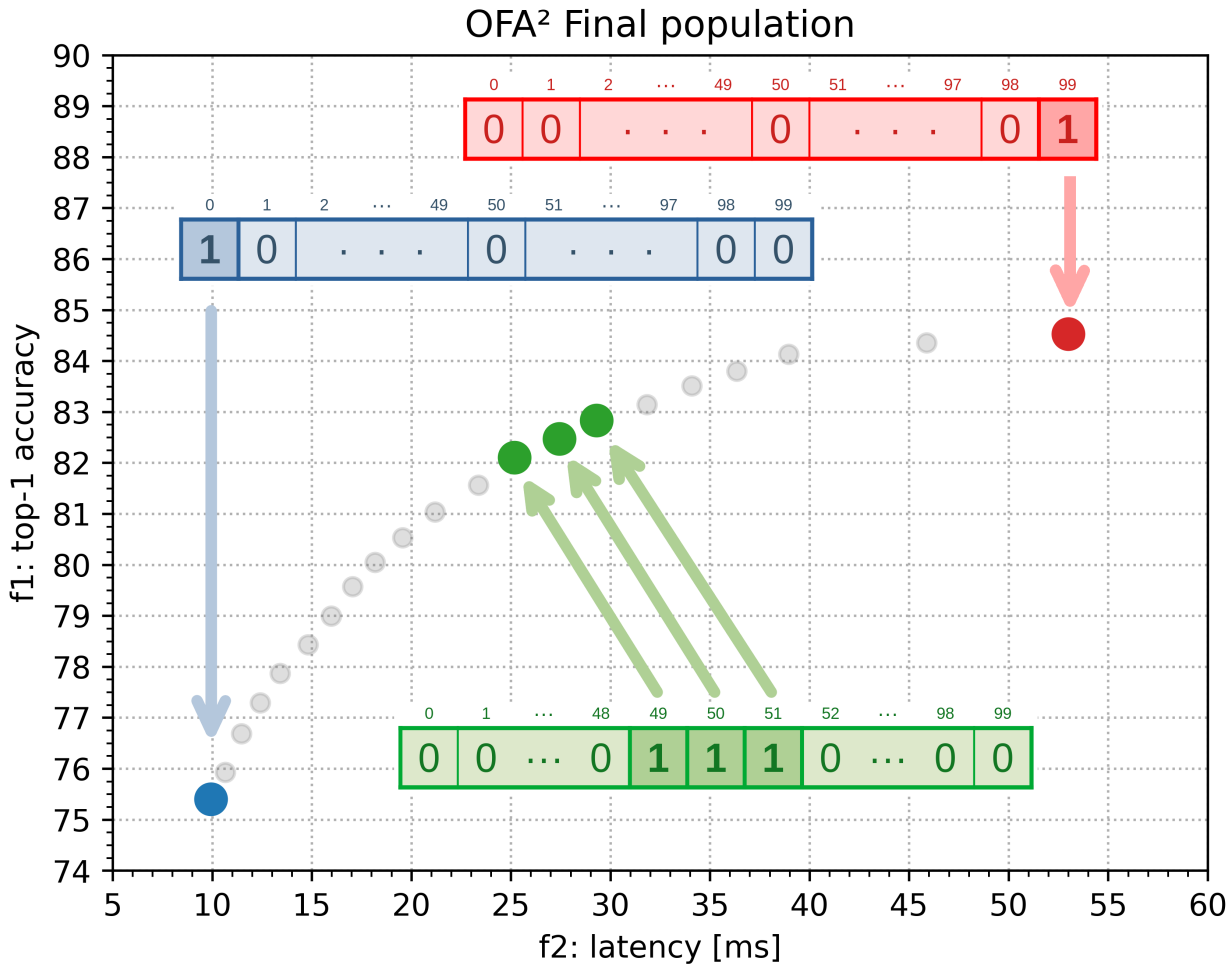
Encoding


OFA² Final population
f1: top-1 accuracy
f2: latency [ms]
0 1 2 ⋯ 49 50 51 ⋯ 97 98 99
0 0 · · · 0 · · · 0 1
0 1 2 ⋯ 49 50 51 ⋯ 97 98 99
1 0 · · · 0 · · · 0 0

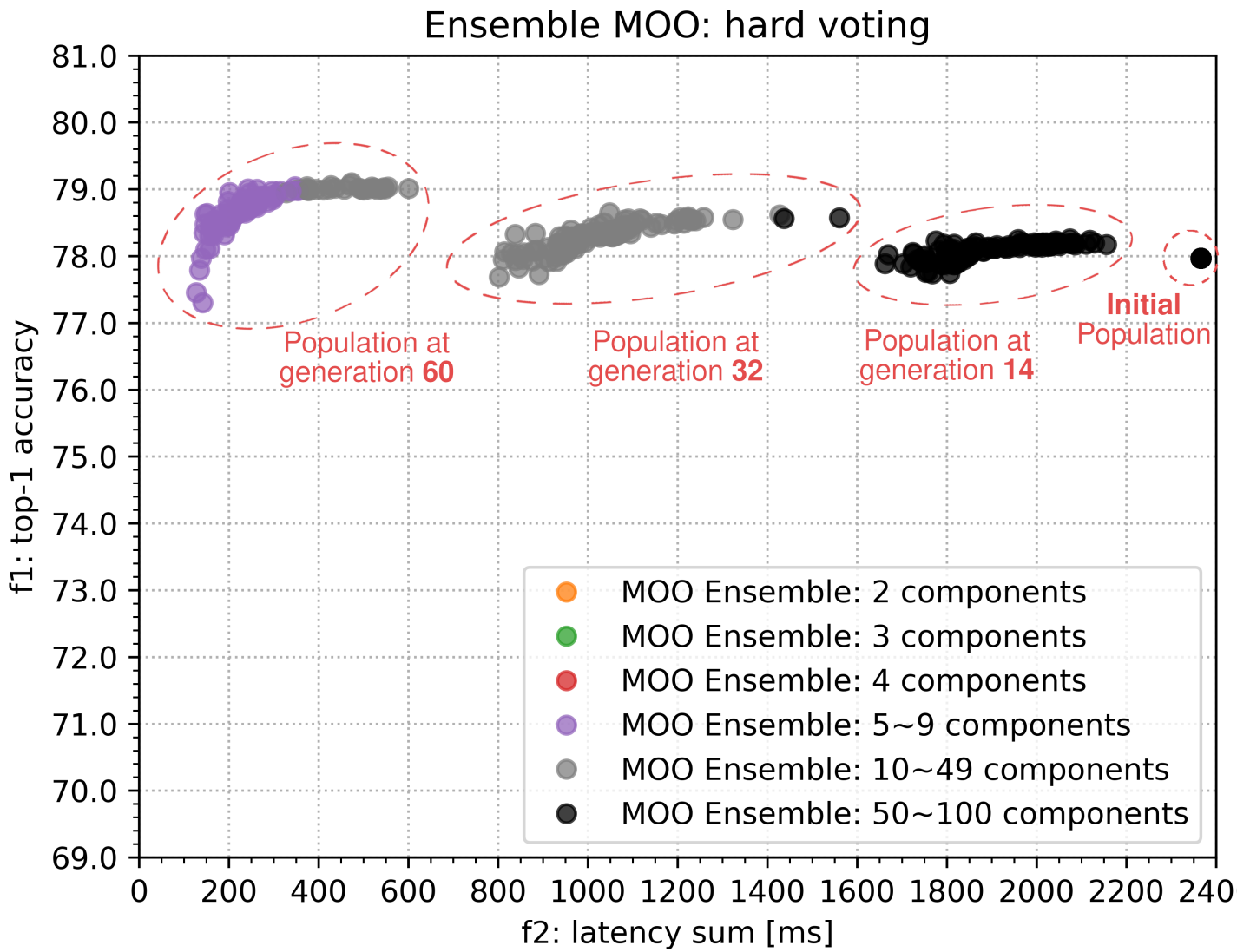
Encoding


OFA² Final population
f1: top-1 accuracy
f2: latency [ms]
0 1 2 ··· 49 50 51 ··· 97 98 99
0 0 · · · 0 · · · 0 1
0 1 2 ··· 49 50 51 ··· 97 98 99
1 0 · · · 0 · · · 0 0

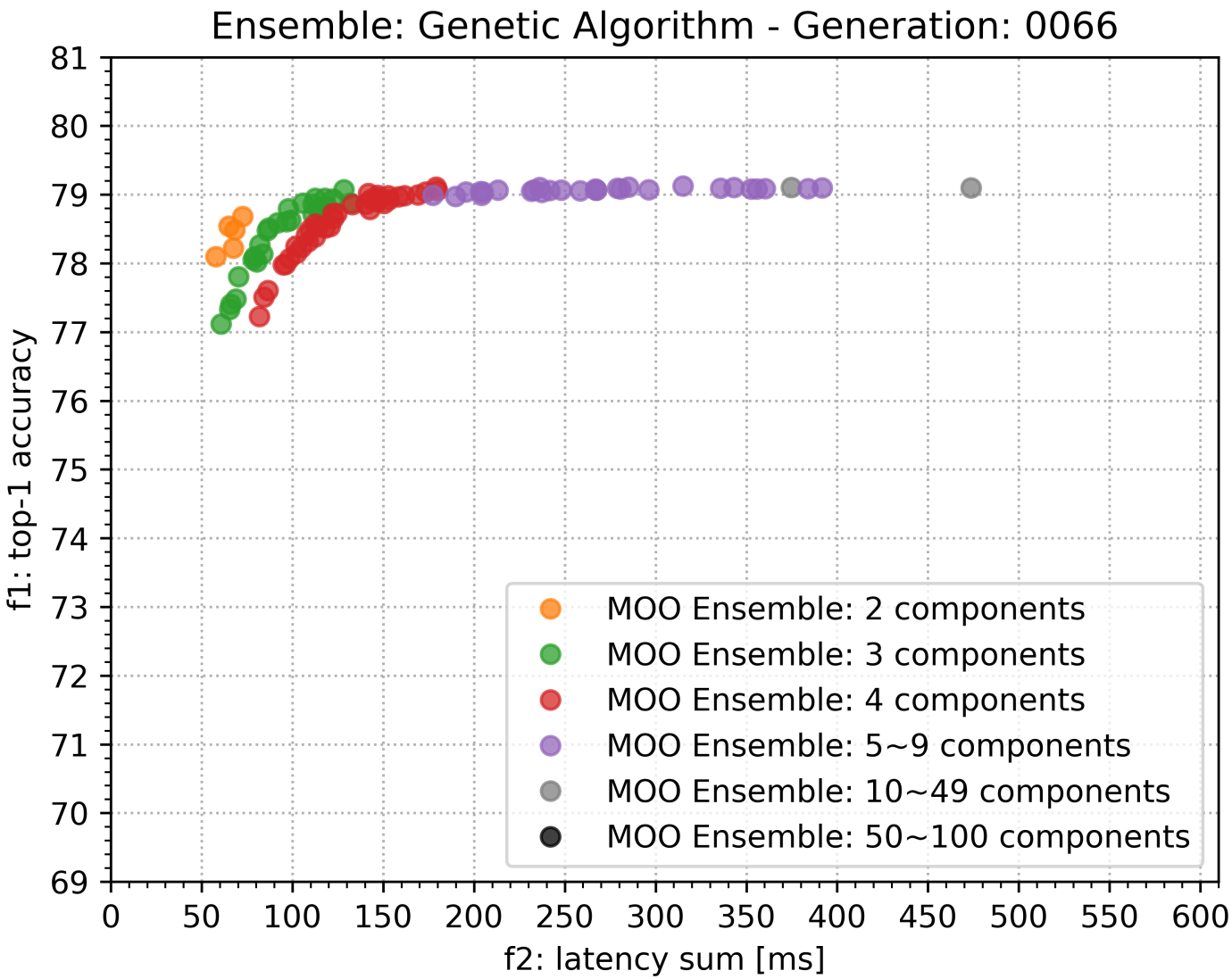
Encoding


OFA² Final population
f1: top-1 accuracy
f2: latency [ms]
0 1 2 ··· 49 50 51 ··· 97 98 99
0 0 · · · 0 · · · 0 1
1 0 · · · 0 · · · 0 0
0 1 ··· 48 49 50 51 52 ··· 98 99
0 0 ··· 0 1 1 1 0 ··· 0 0

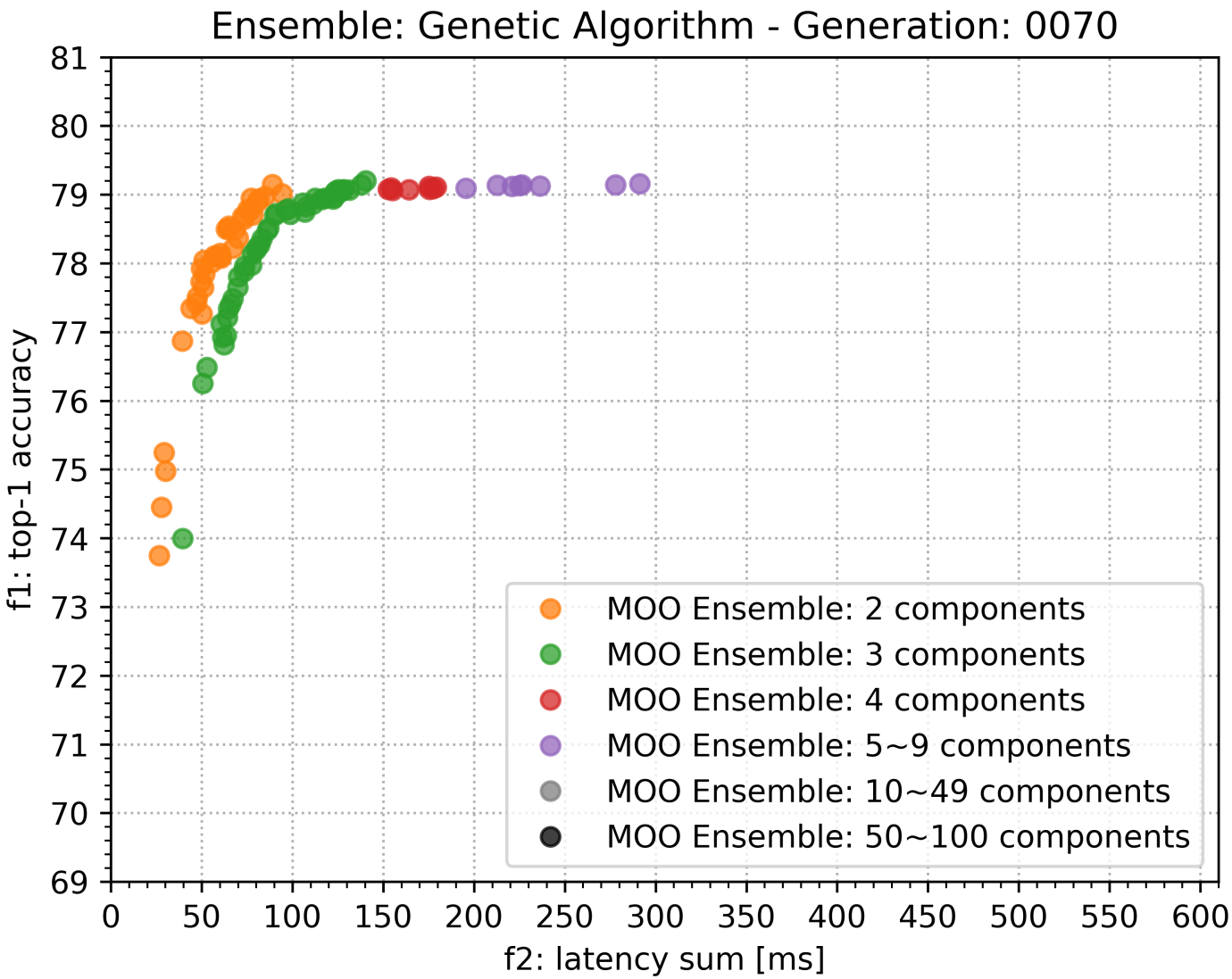
Sum latency: Population progression



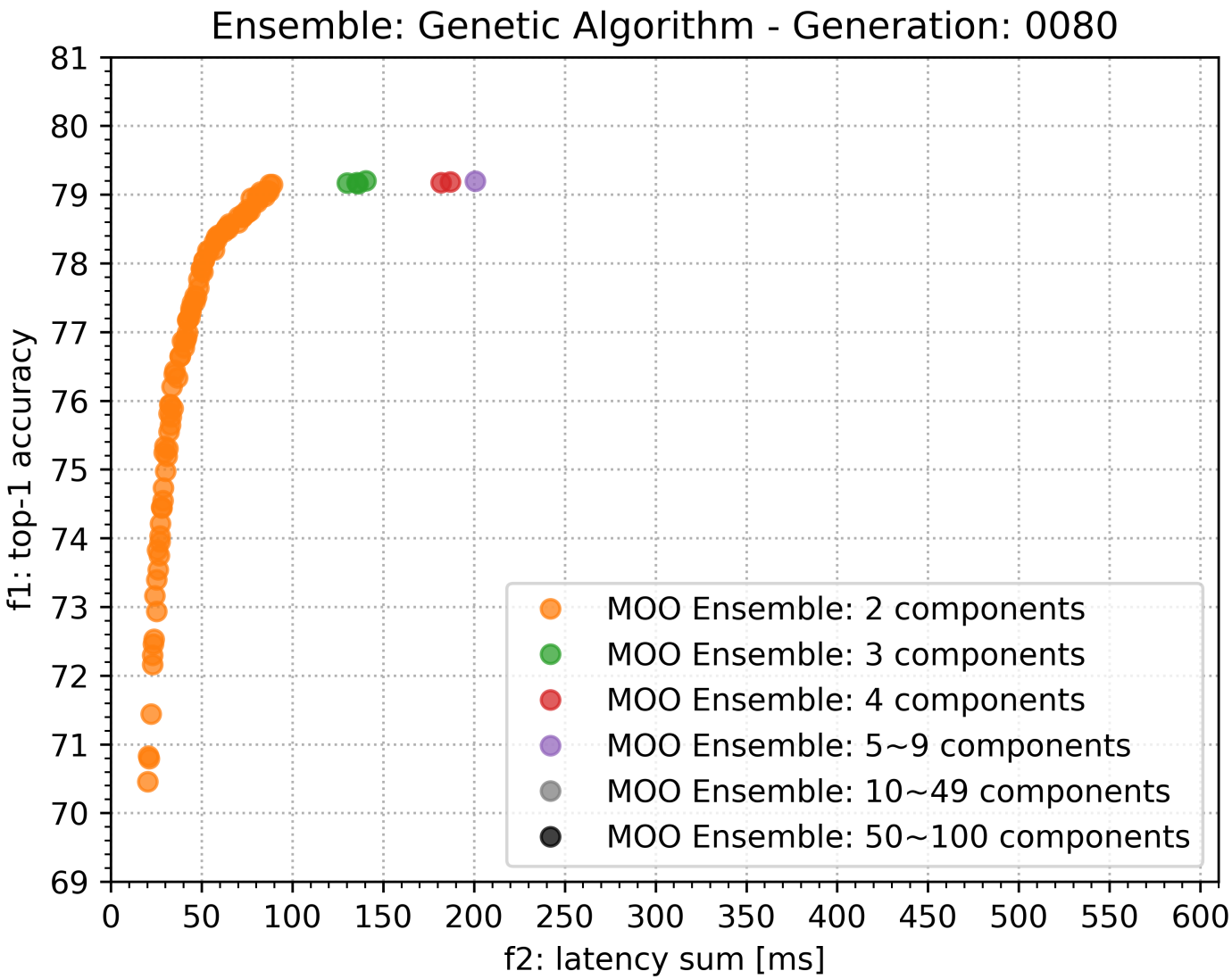
Sum latency: Population progression



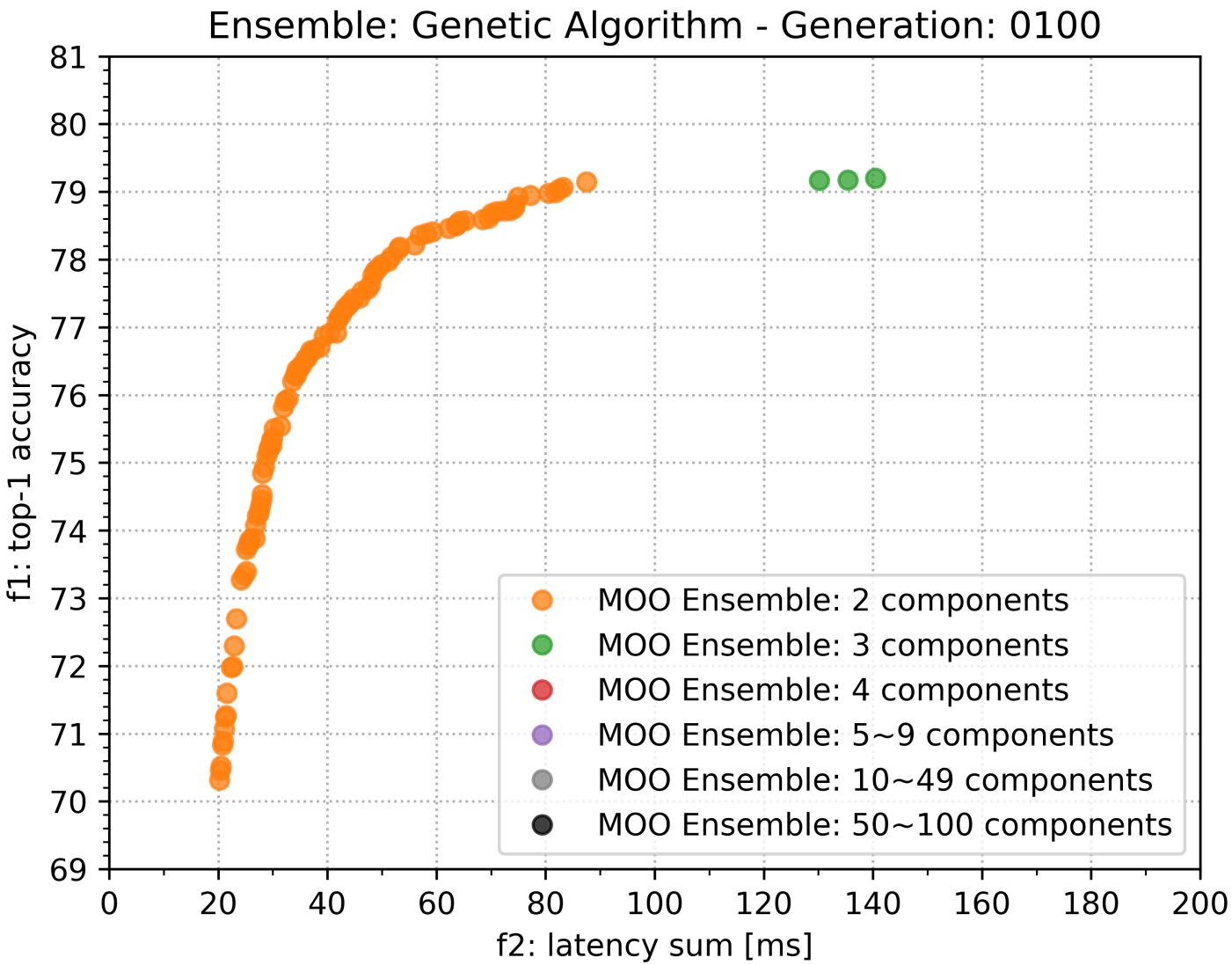
Sum latency: Population progression



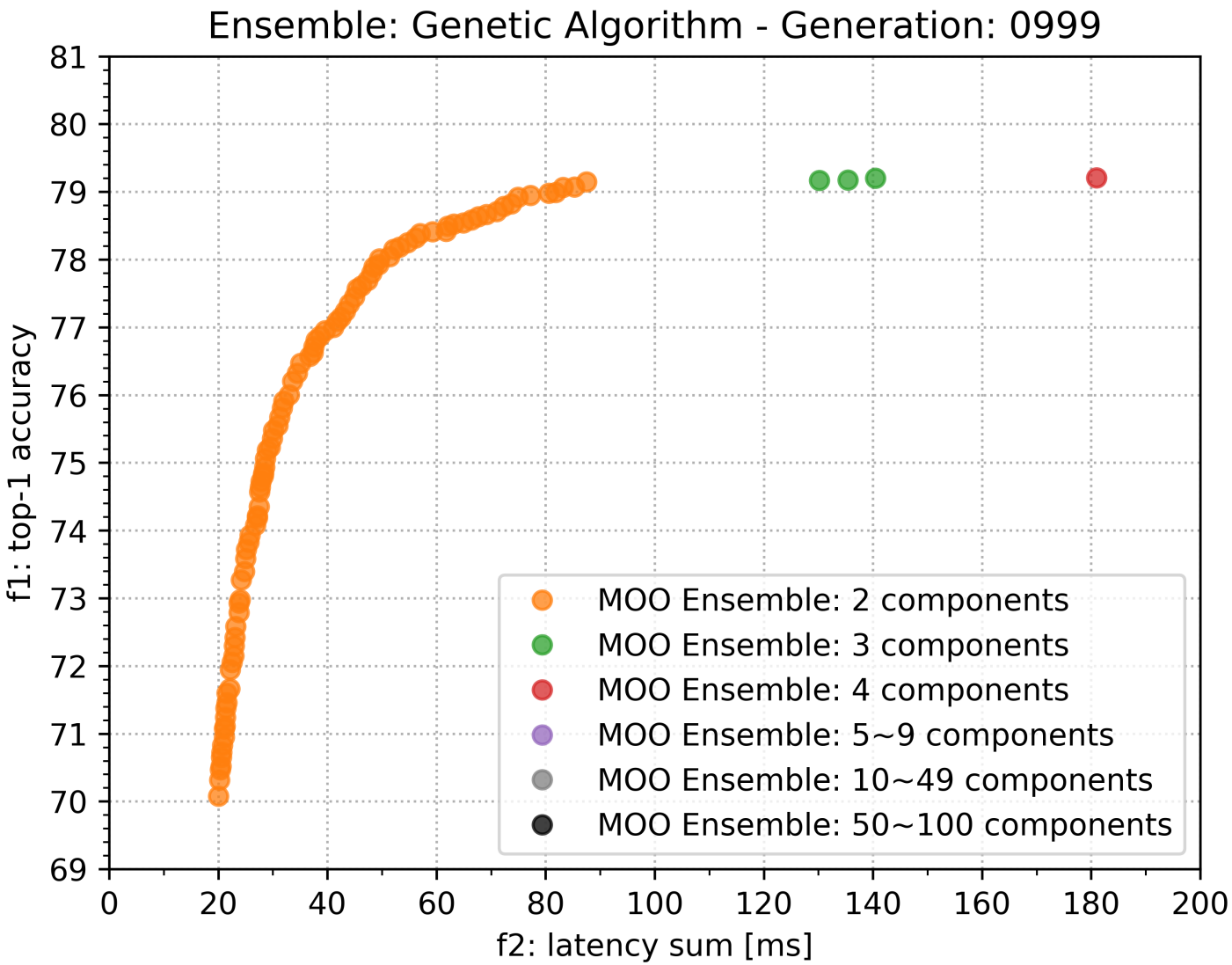
Sum latency: Population progression



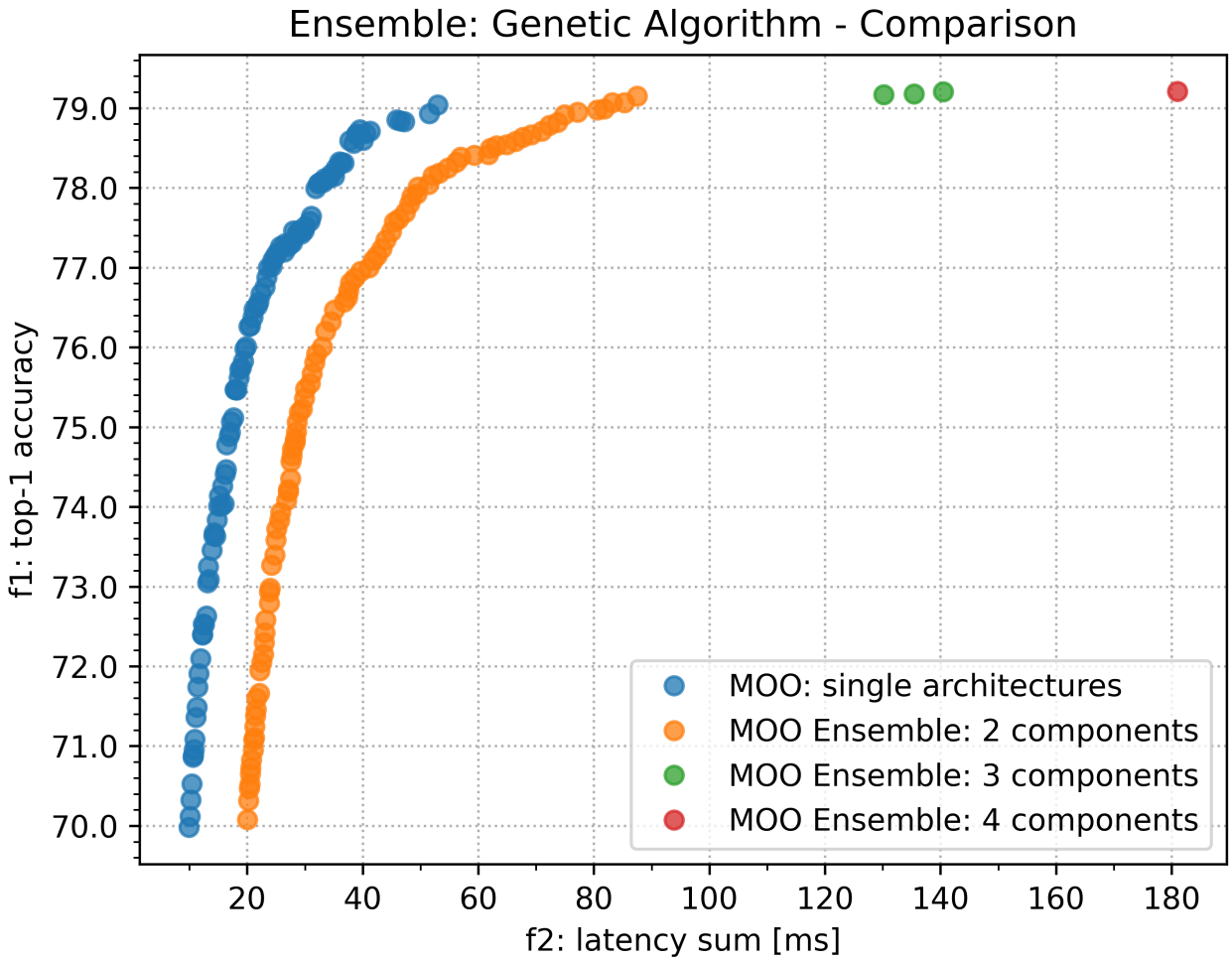
Sum latency: Population progression



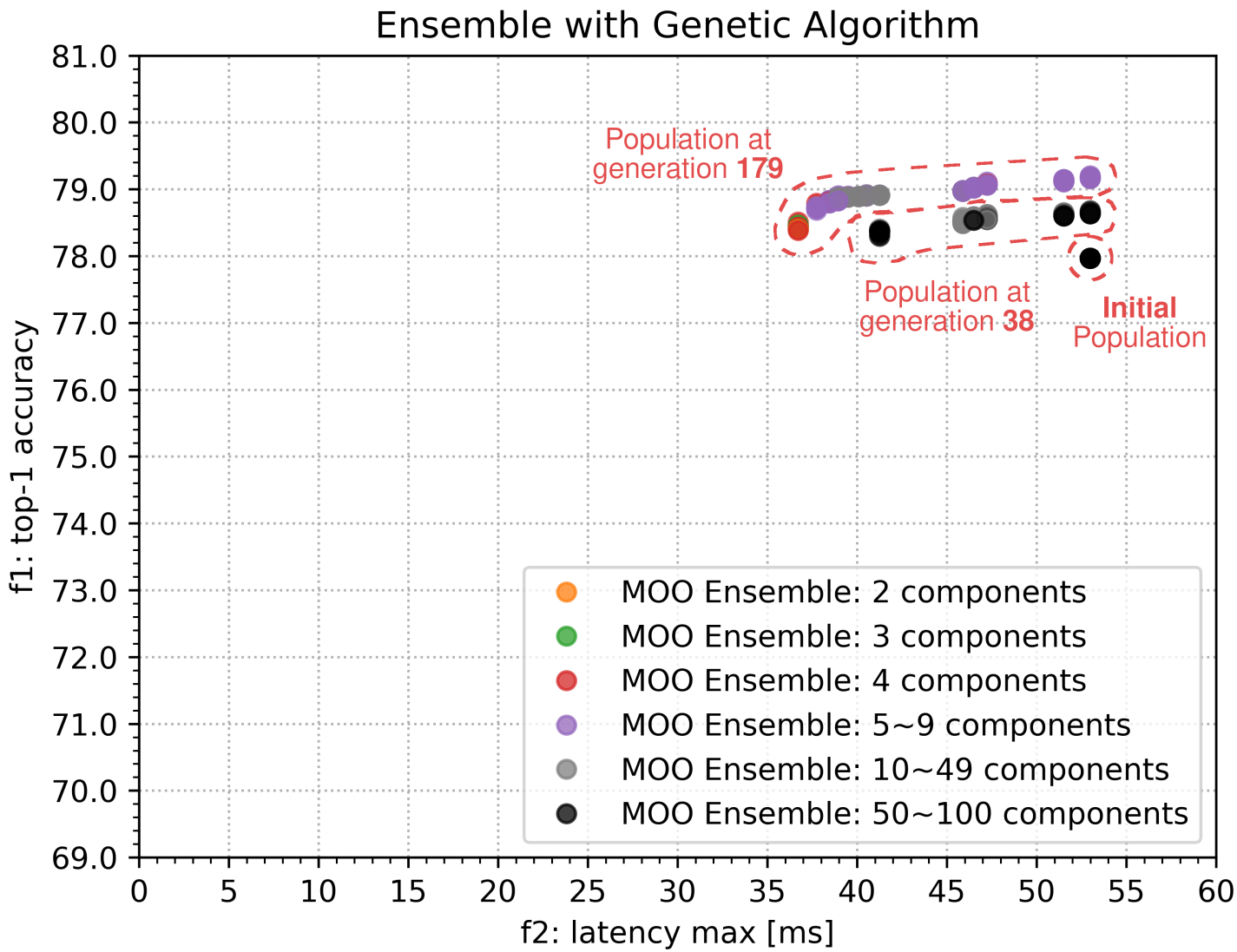
Sum latency: Population progression



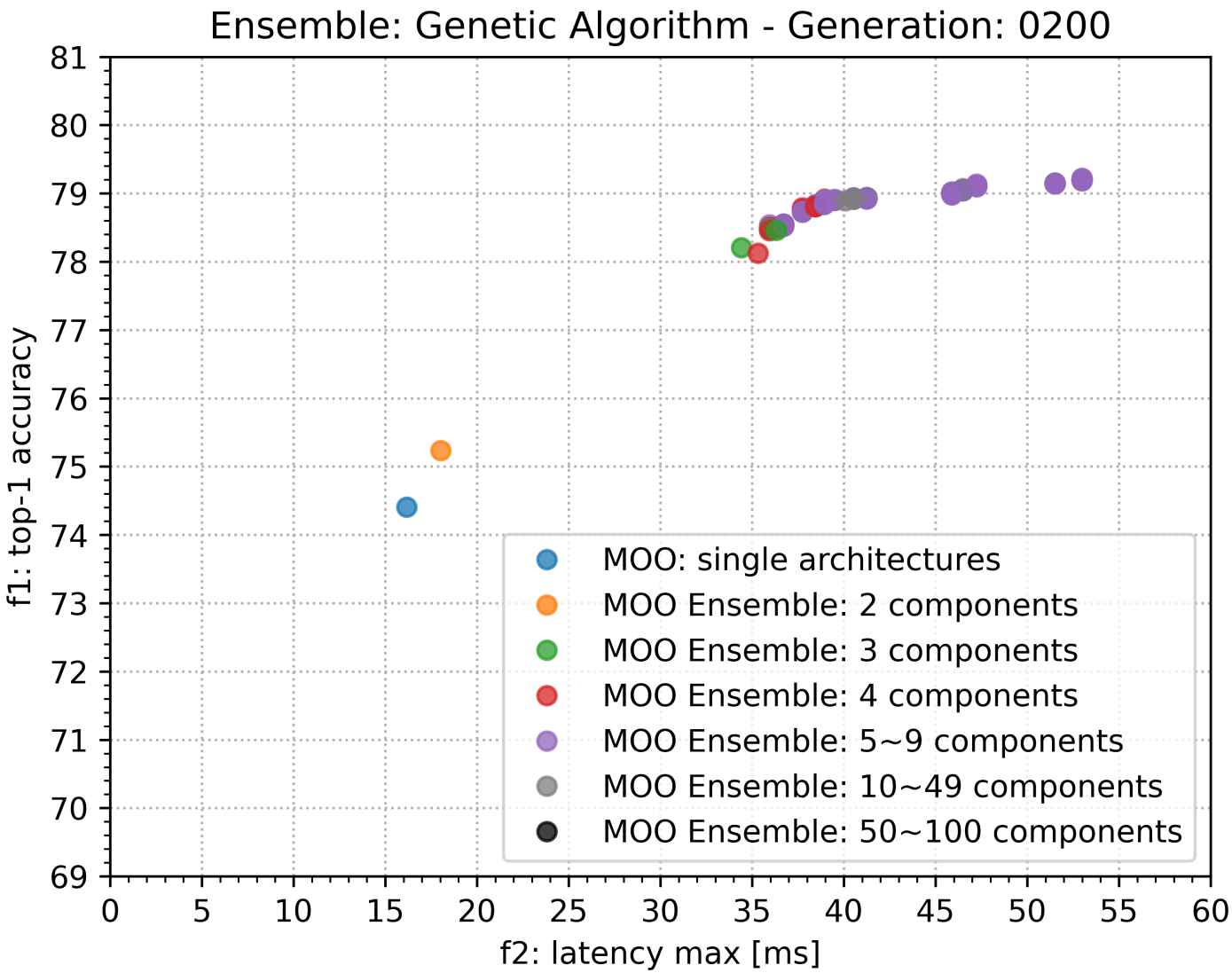
Sum latency: Final population



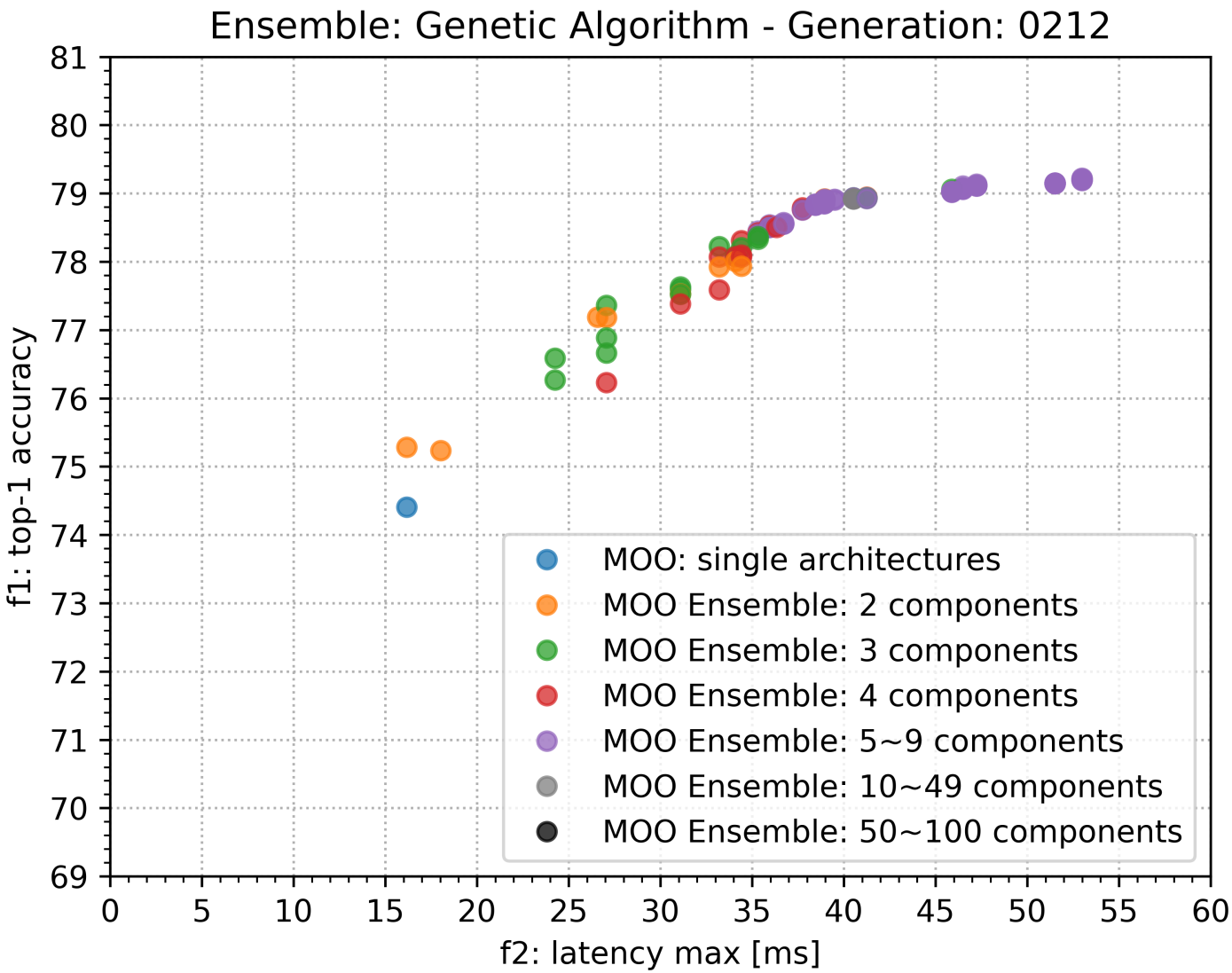
Max latency: Population progression



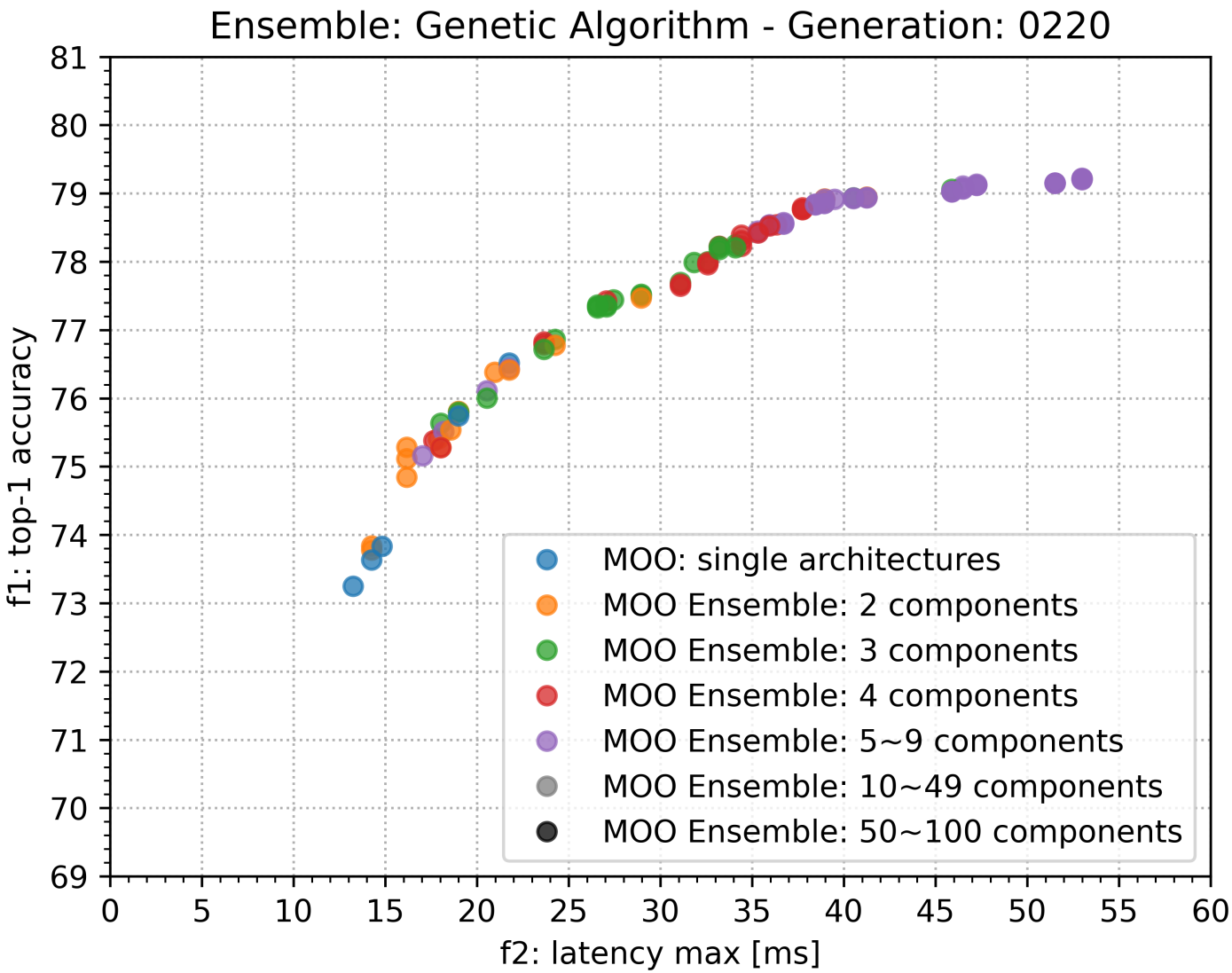
Max latency: Population progression



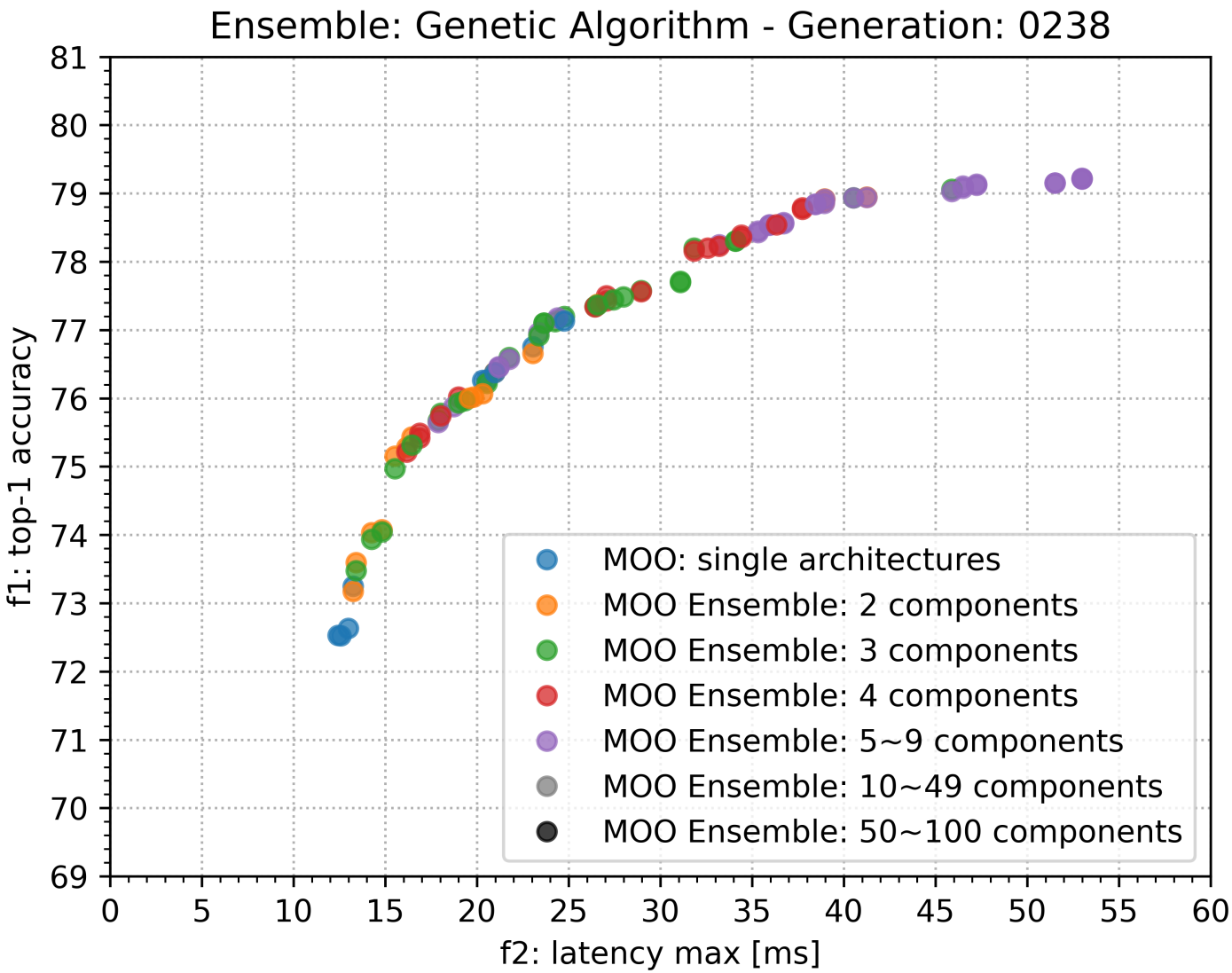
Max latency: Population progression



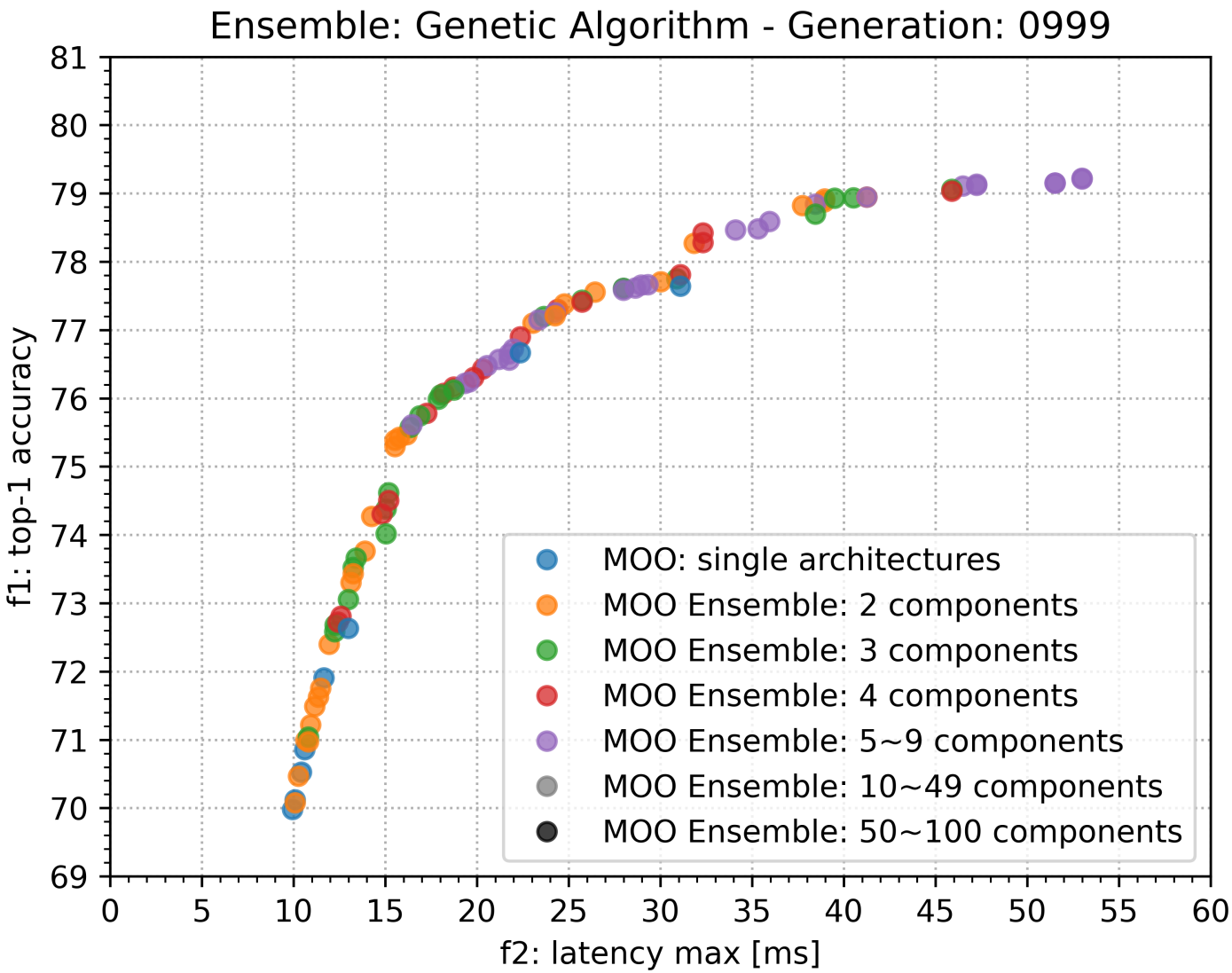
Max latency: Population progression



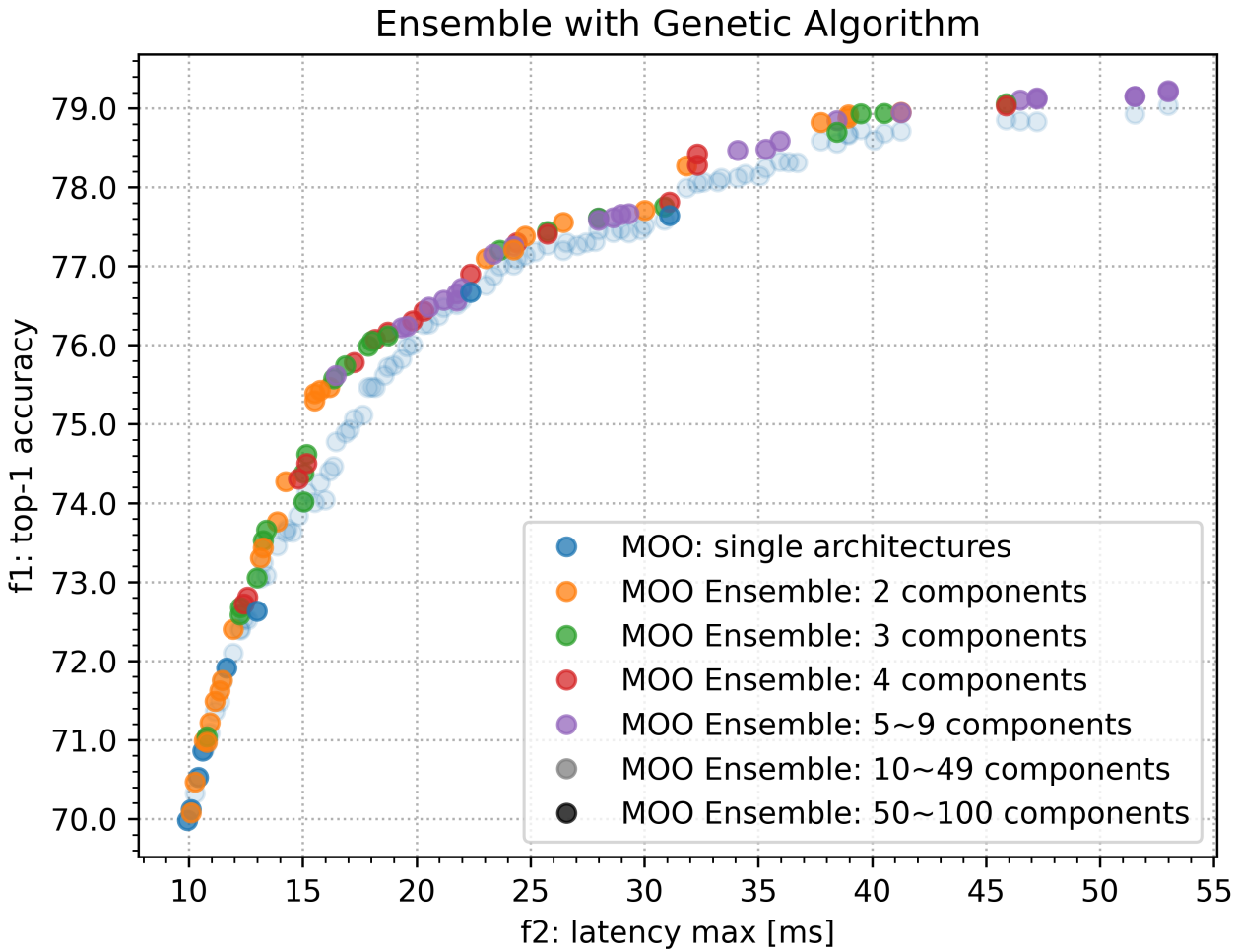
Max latency: Population progression



Max latency: Population progression



Max latency: Final population



Agenda

Future work

- Fine-tuning the search using the real evaluated accuracies of the final population of neural networks obtained with the accuracy predictor.
- Perform the MOO search with all other hardware available.
- Generate accuracy predictor & latency lookup table for more different hardwares

Summary

- OFA² improves the OFA framework by:
  - Finding better architectures with respect to the objectives functions.
  - Find a set of solutions all at once in a single search.
  - Keeps the computational cost roughly the same.
- Proposal of an EMOA encoding for the ensemble formation.
- No clear winner between hard/soft voting strategies for ensembles.
- Committee machine with better accuracies considering the maximum latency (parallel).
- Source code and experiments publicly available.

Agenda

Cai, Han, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. 2020. "Once for All: Train One Network and Specialize It for Efficient Deployment." https://iclr.cc/virtual_2020/poster_HylxE1HKwS.html.

Cai, Han, Ligeng Zhu, and Song Han. 2022. "ProxylessNAS: Direct Neural Architecture Search on Target Task and Hardware." https://openreview.net/forum?id=HylVB3AqYm.

Elsken, Thomas, Jan Hendrik Metzen, and Frank Hutter. 2019. "Neural Architecture Search: A Survey." *Journal of Machine Learning Research* 20 (55): 1–21. http://jmlr.org/papers/v20/18-598.html.

Green, Sam, Craig M. Vineyard, Ryan Helinski, and Çetin Kaya Koç. 2019. "RAPDARTS: Resource-Aware Progressive Differentiable Architecture Search." https://doi.org/10.48550/ARXIV.1911.05704.

Real, Esteban, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V. Le, and Alexey Kurakin. 2017. "Large-Scale Evolution of Image Classifiers." In *Proceedings of the 34th International Conference on Machine Learning*, 2902–11. PMLR. https://proceedings.mlr.press/v70/real17a.html.

Ying, Chris, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, and Frank Hutter. 2019. "NAS-Bench-101: Towards Reproducible Neural Architecture Search." In *Proceedings of the 36th International Conference on Machine Learning*, 7105–14. PMLR. https://proceedings.mlr.press/v97/ying19a.html.